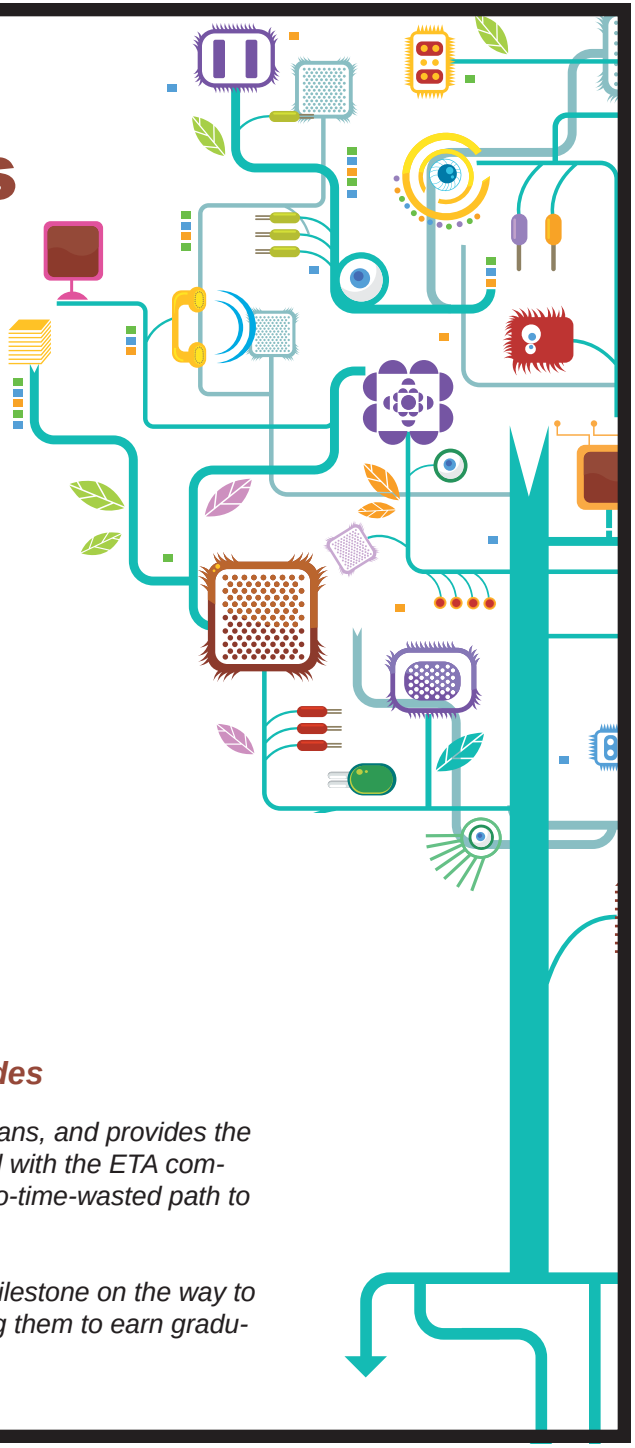


EM3 – Analog Basics

by Karl Eilers



EM3 will cover the following concepts:

- Signal levels, impedances, wiring formats
- Dynamic range, noise, signal to noise ratio, distortion
- Decibels and dB calculations, standard reference levels
- Signal spectrum, bandwidth
- Specific audio, video, RF signals and their characteristics
- Active devices, transistor biasing, amplifier circuits
- Op-amps and other linear IC's
- Oscillators and multivibrators
- Waveshaping circuits
- Filters
- AC-operated power supplies
- Optoelectronics

Give Your Students the Power to Succeed with the Online EM-Series Study Guides

The EM-series of study guides give students the knowledge they need to qualify as entry-level technicians, and provides the necessary foundation for further studies in specialized fields. The study guides are closely coordinated with the ETA competencies and exams. Each study guide builds on the one before it. Together they form an efficient, no-time-wasted path to knowledge and certification.

The basic goal of the study guides is the same as the goal of the EM certifications themselves: each milestone on the way to eventual mastery is useful in and of itself, giving students immediately useful information and permitting them to earn graduated certifications commensurate with their growing knowledge.



Study Guide for ETA-International EM3

ANALOG

by Karl Eilers

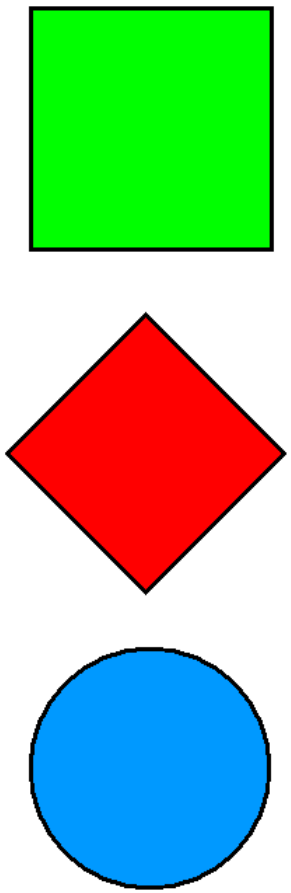
Welcome to the third of the ETA EM6 training modules, ANALOG. It is assumed that you have already completed EM1 and EM2 study guides. In this module, we will discuss analog signals and the electronic equipment that generates and processes them. We will also look at equipment that converts AC from a wall outlet into the DC needed to power most electronic equipment. And we'll look at electronic devices that make light and sense light.

Contents

1. ANALOG SIGNAL BASICS
2. ACTIVE DEVICES
3. TRANSISTOR AMPLIFIERS
4. OPERATIONAL AMPLIFIERS, DC AMPLIFIERS
5. LINEAR INTEGRATED CIRCUITS
6. OSCILLATORS
7. WAVE SHAPING CIRCUITS
8. FILTERS
9. POWER SUPPLIES
10. OPTOELECTRONICS
11. REVIEW

PRACTICE EXAMS

ASPECT RATIO ADJUSTMENT



For many of the illustrations in this study guide, it is important that the dimensions and angles be correct. The green shape should appear as a perfect square, not a rectangle. The red diamond should have 90-degree corners. The blue shape should be a circle, not an oval.

If these shapes are not what they should be, please adjust your display settings before you begin this study guide.

COPYRIGHT NOTICE

All material in this study guide is protected by copyright. When you purchased this study guide, you were given an access code which grants you, and only you, the right to unlimited **personal** use of the material.

You may download a copy at any time and duplicate it as you wish, as long as it's only for your own personal use. You may not share, post on the internet, duplicate for the purpose of sharing, use all or any part in a public presentation or in any way disseminate any part of this material to those who have not purchased the right to use it, unless you first obtain written permission from Karl Eilers or ETA-International.

EM3 Chapter 1
ANALOG SIGNAL BASICS

Objectives of this chapter:

- define “analog” and “digital”
- discuss signal spectrum and equipment bandwidth
- explain decibels and give the formulas for power and voltage ratios
- name common decibel increments and the corresponding power and voltage ratios
- discuss sound pressure level
- discuss standard electrical reference levels
- define “noise floor” and “clipping level”
- give the typical level, dynamic range, spectrum, impedance and format for various audio, baseband video, modulated video and RF signals

Wired and wireless electrical transmissions serve one of two purposes:

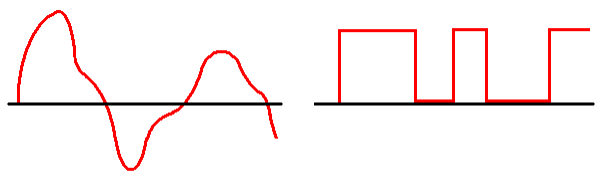
- to move power from one place to another, or
- to move information from one place to another.

An information-bearing transmission is called a **signal**. “Information” is broadly defined; it can be stock market data or a picture of your Great-Aunt Fenestrude.

In this study guide, we're going to focus on information-bearing signals, though a lot of the material also applies to power transmission.

ANALOG VS. DIGITAL

Signals come in two forms: **analog** and **digital**. Digital signals are “on/off” or “high/low”; in other words, they alternate between specific, discrete states. If you've been in a store near closing time and they've flashed the lights, that's a digital signal. The *amount* of light is irrelevant; the signal would mean the same thing if the lights were twice as bright or half as bright.



Analog signals involve an infinite range of levels, and a continuous flow of time. No matter how closely you examine an analog waveform, you won't find any steps or jumps or discontinuities.

A lamp controlled by a switch is digital. A lamp controlled by a dimmer is analog.

EVERYTHING IS AN INTERFACE

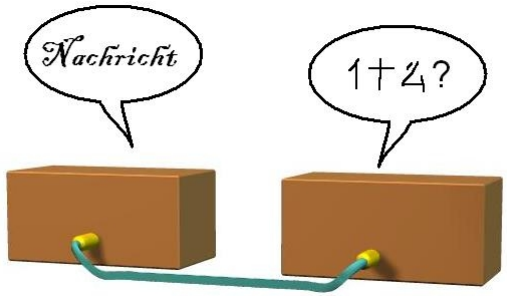
“What we have here is a failure to communicate.”
- Cool Hand Luke, 1967

An **interface** is a junction point where people or electronic circuits transfer information. In the computer world, a “user interface” is something that allows the user to communicate with a computer: monitor, keyboard, mouse. In a radio, the speaker is the user interface between the circuitry and your ears. In your car, the letters PRNDL and the pointer are a user interface between you and the car that tells you what gear the car is in.

An interface has to be properly designed or it won't work. If the radio interface were a light instead of a speaker, if the gearshift indicator were in the trunk rather than in front of the driver, they'd be useless.

The same goes for electronic circuitry. Connections between pieces of equipment, even connections between different parts of the circuit within a piece of equipment, are interfaces. They have to be properly designed or they won't work.

The most basic rule of interfaces is also the most obvious: the circuitry on both sides must speak the same language. You may be familiar with the Biblical story of the Tower of Babel, where each person spoke a language nobody else understood. The electronic equivalent occurs when the signal output of one piece of equipment is connected to the input of another that's designed for a different signal format.



Perhaps the most common example is where someone plugs the output of an MP3 player into a microphone input. The MP3 signal is hundreds of times “hotter” than what the mike input is designed to work with, resulting in massive overload and distortion. There's nothing wrong with either piece of equipment; they just weren't designed to communicate with each other.

INTERCONNECTION STANDARDS

There is no way to design a single interface format that will work in every situation, but we can reduce the Tower of Babel problem by grouping signals by type – audio, radio, video, data, control – and creating a few standard formats for each. If equipment is designed to conform to those standards, we will be able to interconnect it successfully.

The most fundamental aspects of a signal are:

- the signal's **level**, that is, the amount of signal, measured in voltage, current or power. Another word for level is **amplitude**.
- the **dynamic range** over which level may vary; in some cases a 100,000:1 range is normal.
- the signal's frequency **spectrum**, the range of frequencies it contains.
- **impedance**, and whether or not impedances are matched throughout the system.
- **format**; i.e. balanced / unbalanced, shielded, twisted pair, connector types etc.
- **noise and distortion** – how much waveform degradation is permissible.

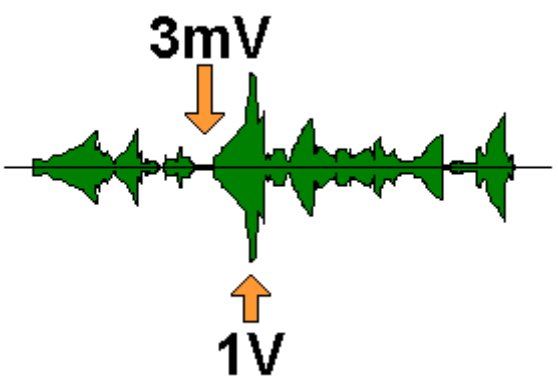
Also, many signals have characteristics specific to that signal type – synchronization pulses in video, for example.

In this chapter, we will look at some of the most common analog signal types. We've already studied impedances and formats in EM2, so this chapter will be devoted to levels and frequencies. As in EM2, I will refer to a piece of electronic equipment as a **unit** for the sake of convenience.

LEVELS

Size matters. If you connect a weak signal to a unit that expects to see a strong signal, it will be inadequate. If you connect a strong signal to a unit built to handle weak signals, it will overload.

In some cases, the signal level throughout a system is held relatively constant. For example, in standard video, instantaneous voltage corresponds to brightness at a particular point on the screen. If the overall voltage is too high, everything on the screen will be overbright and washed out; if it's too low, everything will be dark. So, in video systems, a lot of effort goes into ensuring that signals are always exactly the same level everywhere.

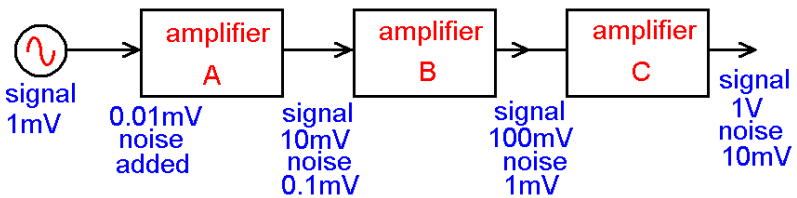


In audio, on the other hand, if the loudest parts of a piece of music correspond to a voltage of 1V, the softest parts may be only 3mV. And if those 3mV parts are to sound good, they have to be essentially free of background noise, which means, for top quality audio, noise can be no higher than 30uV. That's a 30,000:1 range!

In the EM2 chapter on cabling, we studied **signal-to-noise ratio** or **S/N**, the difference between the levels of a desired signal and the undesired noise contaminating it. Typically, for an analog signal to be useful, noise must be very much lower than signal.

THERMAL NOISE

There is a constant, unavoidable low-level noise present in any circuit due to thermal agitation of atoms and molecules. Even when there is no signal, the system itself is always generating a "signal" consisting of noise. The level of this noise dictates the lowest level that can ever exist in the system. It's called the **noise floor**.



In the illustration above, the input signal is 1mVAC, and we're going to pretend it's completely noise-free. (There's no such thing.) At the input to amplifier A, 0.01mV of amplifier noise is added. As the signal passes through amplifier A, and goes on to amplifier B and C, the noise is amplified right along with the signal. It is important to understand that, once noise is added to a signal, other units "downstream" in the chain cannot tell that the noise doesn't belong there. Once noise gets into the signal, from then on it's part of the signal and usually can't be removed. Also, if the signal level is ever permitted to drop below the noise floor, the signal will be lost. (That's not a hard-and-fast rule; sometimes it's possible to retrieve a signal that's buried in the noise.) Ideally, the noise floor will be established by the noise in the original signal and the noise in the first amplifier. After that, the level should be high enough that noise added by "downstream" equipment is insignificant.

Thermal noise is determined by impedance, bandwidth and temperature. At room temperature, a 60Ω resistance will create 1nV (10^{-9}V) of noise over a bandwidth of 1Hz. Noise varies with the square root of impedance and the square root of bandwidth. Quadruple impedance to 240Ω or bandwidth to 4Hz, and the noise doubles. A bandwidth of 10kHz, 10^4 Hz, multiplies the noise by the square root of 10^4, or 10^2 times; 100nV for a 60Ω source. That's pretty close to 50Ω, which is handy for RF noise calculations because most RF cables and equipment are 50Ω.

Does this mean that a 50Ω resistor creates 100nV of noise between 1mHz and 1.01mHz, just sitting there all by itself? Yes, it does. That's the main reason there's a limit to how weak a signal a radio can pick up. No matter how quiet the radio is, the 50Ω antenna impedance generates noise.

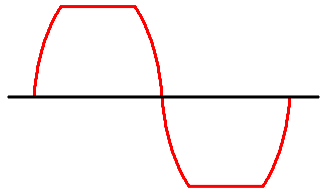
Can't we get rid of the noise by designing a very low-impedance system? That is often helpful, but the places where signal is weakest – at the antenna input, for example – tend to be places where lower impedance also means lower level, so we don't gain anything. There's always a certain signal power and a certain noise power. Changing impedance changes the ratio of voltage to current, but doesn't change the ratio of signal to noise.

The only remedy, when putting together a system, is to:

- keep the signal and noise levels in mind at each point in the system
- set the levels throughout so as to insure that the signal never drops too low
- know how to recognize ahead of time, and be realistic about, a situation where the signal is going to be so weak that it would be impossible for the system to function properly
- know how to change the system design (e.g. change the antenna type) to achieve the necessary S/N ratio

CLIPPING LEVEL

At the other end of the scale from the noise floor, there is a voltage or power **ceiling** – a limit to the maximum signal level a circuit can handle without overload. Exceed this level, and the waveform peaks will be clipped off, which is why this level is also called the **clipping** level. This doesn't always ruin the signal, and in some cases, clipping is necessary. But it always alters the signal, and in most cases, signal alteration is something we want to avoid.

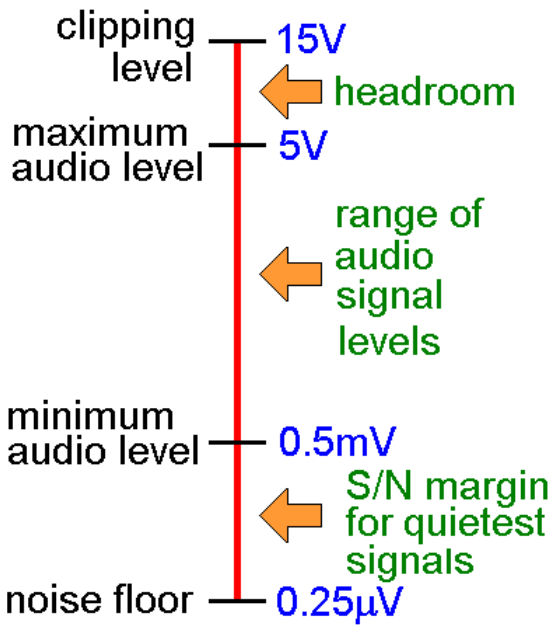


DYNAMIC RANGE

Many signals vary over a range of levels, often approaching both the noise floor and the clipping ceiling. For example, the loudest parts of a symphony may represent 10,000 times as much power as the softest parts. The span of level variations in a signal is its **dynamic range**.

There is a difference between the dynamic range of a *signal* and the dynamic range of a signal *channel*. The dynamic range of a signal is defined by the highest and lowest levels the signal actually reaches. The dynamic range of a signal channel is defined by the highest and lowest levels a signal could reach, the highest and lowest levels it could properly handle, which is to say the range between the noise floor and the clipping ceiling. In a low-quality audio system like a telephone or dictation recorder, the system is adequate if it has a 30:1 (voltage) dynamic range. High quality music requires a 30,000:1 range. Video should be several hundred to one or better.

When a microphone is inside a kick drum, it will put out peaks of perhaps 5 volts. The circuitry it's connected to should be able to handle a higher voltage than this, so as to provide some safety margin – 15 volts, say. This margin, the difference between the clipping level and the highest- level signal that is actually likely to occur, is called **headroom**.



On the other hand, when the microphone is picking up a choir singing softly, its output might be 0.5mV. The circuitry it's connected to should have a low enough noise floor not to contaminate the signal, which means it must be a lot lower than 0.5mV.

BTW, just to head off some confusion, a microphone that uses magnets and a coil of wire to generate a signal – a speaker in reverse – is called a “dynamic” microphone. That name has almost no connection to dynamic range.

A low-impedance dynamic microphone generates around 0.25 microvolt of noise all by itself. The best amplifiers are significantly quieter, so the major source of noise is the microphone, which defines the noise floor. So, let's figure this out: we're dealing with a maximum level of 5×10^{0} volt and a noise floor of about 3×10^{-6} volt, or a dynamic range of more than a million to one.

Yikes! We've got to handle signal levels that vary by thousands to one, and headroom above that, and a noise floor below that, and this all gets pretty number-intensive. Fortunately, there's a way to simplify things. It's called the **decibel**, abbreviated **dB** – small d, capital B.

DECIBELS

Decibels are a logarithmically-based way of stating level differences. Today, it is applied to all kinds of signals, but it was first developed for use with audio signals; that is, sounds.

When we're dealing with subjective impressions, it's very hard to establish objective benchmarks. One person's “loud” is another person's “quiet.” A recording of a person screaming will seem louder than a recording of a person whispering, even if the actual levels are the same.

However, people do seem to agree on loudness *changes*. If you play a recording, and then play it again while turning it up, most people will agree on when it's "twice as loud" as the first time you played it. If you start turning it down, most people will agree on when it's "half as loud." From this fact, telephone engineers derived the **bel.**

Definition: a Bel is a subjective doubling or halving of loudness.

The abbreviation is **B**, capitalized. And yes, it's named after Alexander Graham Bell.

It turns out that a subjective doubling or halving of loudness very roughly corresponds to raising or lowering the electrical power level by a factor of ten. That is, if you boost the power by 10 times, it will seem twice as loud; if you cut it to $1/10^{th}$ of its original power level, it will seem half as loud. It may seem strange that the human ear requires ten times the power to perceive two times the loudness, but the ear is built to handle an immense loudness range. The difference between the quietest sound you can hear and the loudest you can tolerate is about 120dB, twelve doublings of subjective loudness, or an acoustic power range of 10^{12}-to-1. That's a trillion to one. No surprise, then, that the ear isn't very sensitive to small power changes. So:

Definition: electrically, a Bel is multiplication or division of a power level by a factor of 10.

This gives us a simple way to attach numbers to sound level changes by measuring logarithmically. Officially, the bel is defined as a 10:1 power ratio. One Bel up, ten times the power, double the subjective loudness. One Bel down, one tenth the power, half the subjective loudness.

Like the Farad, a Bel is too large and crude an increment to be generally useful, so it's divided into smaller units. A bel is divided into 10 smaller parts, called **decibels**, abbreviated **dB** – small d, capital B. Memorize this:

10dB is a subjective doubling or halving of loudness, and/or multiplication or division of a power level by a factor of 10. +10dB = 10X power, -10dB = $1/10^{th}$ power.

Smaller dB increments involve raising or lowering the power level by a factor smaller than 10.

The decibel scale is a *linear* representation of a *logarithmic* quantity. Recall from EM2 that, on a log scale, every 10:1 span is a **cycle**. For example, a scale that covers the range between 10 and 10,000 is a 3-cycle scale:

cycle 1:	**10-100**
cycle 2:	**100-1,000**
cycle 3:	**1,000-10,000**

We also use the word **decade**, which in math doesn't mean a ten-year time span but a 10:1 quantity span. Every decade increase in power is 10dB. If we start with 1mW of power and then boost it to 10mW, then 100mW, then 1W, then 10W, we are boosting it by 10dB, 20dB, 30dB and 40dB respectively. If we start with 1mW and cut it to 100μW, then 10μW, then 1μW, then 100nW, we are lowering it by 10dB, 20dB, 30dB and 40dB. Thus: a linear progression of numbers to represent a logarithmic progression of power levels.



This gives us an easy way to handle and compare huge differences in level. We could talk in terms of a power ratio of ten million-to-one, or we can just say “70dB.”

One of the biggest misunderstandings about decibels is the idea that they represent absolute power levels. They do not, at least by themselves. A decibel number represents a *change* from one level to another. An increase in level corresponds to a positive decibel number, like +6dB. A decrease in level corresponds to a negative decibel number, like -10dB. If a level stays the same, that represents an 0dB change.

This is true whatever the actual levels happen to be. If we start at 1mW and increase to 10mW, that's +10dB. If we start at 100W and increase to 1,000W, that's also +10dB. Decibels are *ratios*.

DECIBELS AND POWER CALCULATIONS

The formula used to convert a power ratio to decibels is:

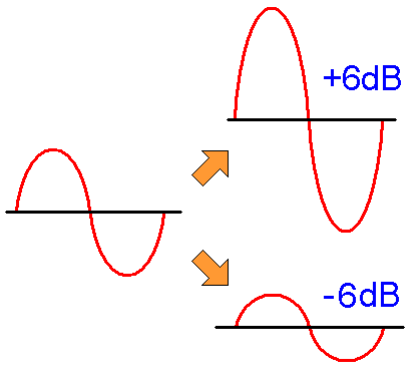
$$dB = 10 \log (P_2 / P_1)$$

where P_1 is the initial power level and P_2 is the power level after the change. That's an easy formula if you have a scientific calculator.

Let's work an example. Suppose we take an existing signal and boost its power 4 times.

$$P_2 / P_1 = 4$$
$$\log 4 = 0.602$$
$$\text{round to } 0.6$$
$$10 \times 0.6 = +6dB$$

So, boosting the power 4x results in a 6dB increase in level, which we would write as +6dB. Dividing the power by 4x would result in a 6dB decrease in level, or -6dB.



Now that you've used the equation, I have to tell you that most techs never bother with it. For one thing, you won't always have a scientific calculator with a log button in your pocket. For another, if you want to go the other way – if you have a dB increment and you want to find the power ratio – you need to do an antilog calculation, which gets a bit involved. Instead, most techs just memorize a few of the most common ratios. The following is also in the review chapter, so you can print it out.

- **1dB = 26% increase or decrease in power**
- **3dB = doubling or halving of power**
- **6dB = 4x or ¼x power**
- **10dB = 10x or $^{1}/_{10}$x power**

Only the 10dB figure is precise. The others are approximations, but they're close enough that the errors are insignificant, even in repeated calculations.

From there, you can get pretty close to any dB ratio by doing repeated calculations. The rule is, **when you *add or subtract* dB's, you *multiply or divide* power levels**. For example, suppose we want to increase a level by 7dB. First, add and/or subtract known dB ratios to get 7dB:

+7dB = +10dB -3dB

... so we first need to go up by 10dB, then down by 3dB.

+10dB = 10x power
-3dB = ½ power
so **7dB = 10 x ½ = 5x power**

Those are easy enough that you can do them in your head, meaning you can get within 1dB of any power ratio without a calculator. In most real-world cases, anything smaller than 1dB doesn't matter. Then, for increases or decreases greater than 10dB, we invoke our knowledge of powers-of-ten:

+20dB = 100x power
+30dB = 1,000x power
+40dB = 10,000x power

...and so on. Note that every 10dB change involves changing an exponent by 1, as in 10^3 mW to 10^4 mW.

DECIBELS AND VOLTAGE CALCULATIONS

Voltage ratios work the same as power ratios, except that a given voltage change results in twice as great a dB change as when we were dealing with power:

- **2dB = 26% increase or decrease in voltage**
- **6dB = doubling or halving of voltage**
- **12dB = 4x or ¼x voltage**
- **20dB = 10x or $^1/_{10}$x voltage**

We will often still want to use 3dB and 10dB, so here are the numbers:

- **3dB = square root if two (about 1.4) times voltage, or square root of ½ (about 0.7) times voltage**
- **10dB = square root of 10 (about 3.2) times or square root of $^1/_{10}$ (about 0.32) times voltage. The fact that 3.2 is the alcohol percentage of naturally fermented beer is purely accidental.**

The formula is:

$$dB = 20 \log (V_2 / V_1)$$

The important thing is that the voltage formula says “20 log...” instead of “10 log...”

The reason that a given voltage change represents twice as many dB as a given power change (for example, 2X power = +3dB but 2X voltage = +6dB) is that power varies with the *square* of voltage. If you double the voltage you double the current, and since power = voltage X current, you quadruple the power. That's doubling power twice, +3dB twice, +6dB.

Decibels give us a way to make useful comparisons between levels that are far apart, perhaps by orders of magnitude, which we'll need to do often. For example, if there's 1 microvolt of signal at a radio's antenna input, and the radio's processing circuitry needs a 1 millivolt signal to work with, we know we need a 60dB amplifier, which corresponds to a 1,000X voltage gain.

EXERCISES

1. Suppose a speaker is being driven by 1 watt of audio. Suppose the volume is turned up so the power level is 20 watts. What dB increment does this represent?
2. In question 1, what will the subjective increase in loudness be?
3. A volume control typically attenuates a signal by 5:1 at mid-position. How many dB's of attenuation is this?
4. Multiplying a voltage 20x is a ____ dB boost.
5. Multiplying power by 8x is a ____ dB boost.
6. a 37dB gain is what voltage ratio?

Answers

1. 20 = 2 x 10. Multiplying power by 10x = +10dB. Multiplying power by 2x = +3dB. So 20x power = 10 + 3 = **+13dB**. Note the use of the plus sign, indicating a positive change in level.
2. 10dB is roughly a doubling of perceived loudness. 3dB is a small loudness change. So, the sound should seem to be **a bit more than twice as loud**.
3. 15dB is 6 + 6 + 3dB, which is 2 x 2 x 1.4V. That's about **5.6** times the original voltage.
4. 10x voltage = +20dB, 2x voltage = +6db, so the total is **+26dB**.
5. 8 is 2 x 2 x 2, which is 3 + 3 +3dB, or **+9dB**.
6. +37 is 20 + 20 -3. +20dB corresponds to 10X voltage. -3dB is 0.7x voltage. So +37dB is 10 x 10 x 0.7 times the voltage, or **70x**. Someone who does these calculations all the time (like you, in the future) would automatically know that +40dB is 100X. 3dB less is 100X times the square root of ½, 100 X 0.7.

As noted above, decibels were originally invented for audio, but the concept proved so useful that signals of all kinds are routinely calibrated in decibels. When I had a land line internet connection, the level was too high, so the tech inserted a 3dB attenuator. DBs are used all over.

DECIBELS AND ABSOLUTE LEVELS

As stated earlier, a decibel is a change between levels, *not* an absolute level. However, we can choose an absolute level as a **reference**, which then becomes “0dB” for our purposes. From there, we can use dB's to indicate levels above and below that reference.

For example, a person who says a signal level is “+10dB” is talking nonsense, but a person who says a signal level is “+10dB relative to 1 volt” is properly identifying an absolute level of about 3.2 volts.

In many cases, the reference level is so well known to the techs who use it that they don't mention it. In professional audio work, when someone says “+4dB,” everybody knows the reference level is 0.78V, which makes “+4dB” about 1.23V. But that implied reference level must be there. (Why would anybody choose 0.78V as a reference level? Read on.)

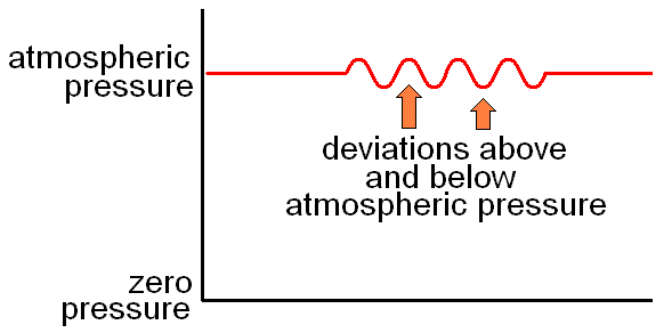
ACOUSTICAL REFERENCE LEVELS

Most students of this study guide aren't likely to become sound engineers. However, the concepts may be easier to grasp if we talk about sounds, which you can sense, rather

than radio signals, which you can't. The way sounds radiate from a sound source in air is similar to the way radio signals radiate from an antenna, so the concepts are worth some study.

Sound Pressure Level, SPL

SPL is the measure of deviation from atmospheric pressure. It's an alternating quantity – sound waves alternately add to and subtract from atmospheric pressure – and it's measured using the RMS method. It's the acoustical equivalent of a small AC voltage riding on a large DC voltage.



The illustration above shows pressure deviations associated with an extraordinarily loud sound. Normal conversation involves about 0.000000000002% variation in atmospheric pressure.

The quietest sound an average person can hear was chosen as the acoustical reference of **0dB SPL**. There are quieter sound levels than this but you can't hear them because they're below the **threshold of hearing**. Thus anything you can hear corresponds to a positive SPL. Conversation, for example, tends to be around 60dB SPL, while a jackhammer five feet away may reach 100dB SPL. In the case of sound pressure levels, we omit the "+" sign because it's safe to assume that all the SPL's we normally deal with are positive.

There are places that are quieter than 0dB SPL – that is, their SPL numbers are negative. The world record, according to Guinness, is a laboratory in Minneapolis that measures -9.4dB SPL. That's an extremely low level and difficult to achieve.

If the source of a sound is a point that radiates sound in all directions – if it's **omnidirectional** – SPL falls 6dB with every doubling of distance from the source. The same is true of an RF antenna that radiates in all directions. It's also true of most sources that are not omnidirectional, but there are special circumstances that can invalidate the rule. So we'll stick with omnidirectional sources for the moment.

SPL is the most commonly used acoustical reference, but there are two others that I'll just mention in passing.

Sound Power Level (Wattage), SWL

SWL is the total sound power emitted from an acoustic source, measured in watts. 0dB SWL is 10^{-12} watt.

Sound Intensity Level, SIL

SIL is the measure of sound power per square meter, the energy that passes through a square meter at a particular distance from the sound source. It's the combination of sound pressure and air particle velocity. You'll recognize the similarity to electrical power, which is the combination of pressure (voltage) and movement (current). SIL falls 12dB with every doubling of distance from the sound source. The same is true of the intensity of electrical energy radiated from an omnidirectional antenna.

ELECTRICAL REFERENCE LEVELS

Sound is pretty much limited to 20kHz, but electrical signals can of course contain any frequency or mix of frequencies. There are multiple electrical reference levels; we use whichever makes most sense for a particular kind of signal.

With electrical signals, unlike SPL, the reference identifies not a minimum level but something like **the average, normal level for that kind of signal**. This means we will routinely encounter levels both above and below the reference, so now we cannot omit the "+" sign before a higher level as we did with SPL.

For each standard, a letter is added after "dB" to indicate the reference. For example, one volt AC is often a handy reference. Any number referred to one volt would be stated in **dBV**. From there, we go up or down:

- "+ so many dBV" is a level above the reference. For example, **+6dBV** is **2 volts**.
- "- so many dBV" is a level below the reference. For example, **-6dBV** is **½ volt**.
- **0dBV** is a level equal to the reference, which of course means it's **1 volt**.

It's the letter after "dB" that tells you you're dealing with an absolute level rather than a change up or down.

Different situations call for different references. Voltage-related references are:

- **dBV**, decibels relative to one volt.
- **dBmV**, decibels relative to 1 millivolt.
- **dBµV**, decibels relative to 1 microvolt. These lower references might identify levels at an antenna input, which typically are in the microvolt to millivolt range.

Power-related references are:

- **dBm** – decibels relative to one milliwatt, the traditional reference for telephone land lines.
- **dBW**, decibels relative to 1 watt.
- **dBμW**, decibels relative to 1 microwatt. When the reference is a power level, the impedance of the circuit determines the actual voltage and current, according to the Ohm's and Watt's laws you learned way back in EM1. For example, in a 600Ω circuit, 0dBm corresponds to a voltage of 0.78V. In a 50Ω circuit, it's a bit over 0.22V. If you work out the Ohm's and Watt's law formulas you'll see why.

Two references used only in professional audio:

- **dBu** – decibels relative to 0.78 volt.
- **+4dBu** – 4dB above 0dBu, about 1.23 volt. This level reads "0" on a traditional VU meter.

The dBu doesn't make sense unless you know the history. Professional audio technology is an outgrowth of land line telephone technology, which was based on a standardized, matched impedance, usually 600Ω. The the standard signal reference level was 1mW or 0dBm which, in a 600Ω circuit, corresponds to a voltage of 0.78V.

Beginning in the 1970s, audio technology moved away from matched impedances. That meant that the 1mW standard was obsolete, and the 0dBm reference was meaningless. To avoid the interconnection problems that would have occurred if levels had been changed, the existing standard was simply redefined. The pro audio reference level today is 0dBu, which is 0.78V, the actual voltage we were dealing with in the 600Ω days, but not tied to impedance or power any longer. Today, about the only place where impedances are matched is in the land-line telephone system, where lines are so long they could have reflection problems even at audio frequencies.

To review:

- dB's are linear numbers that correspond to logarithmic power or voltage changes
- 10dB = 10X power change
- 20dB = 10X voltage change
- dB's may identify an absolute level when tied to a reference power or voltage. The reference is usually identified as "0dB" with a letter or several following to identify the reference.

REFERENCE LEVELS VS. ACTUAL LEVELS

A reference level is not necessarily the actual voltage or power level normally encountered. For example, in consumer audio equipment, the *reference* is 0dBV, but the *actual voltage* is usually on the order of -10dBV OR 0.32V.

EXERCISES

1. A +6dBV signal is _____ volts.
2. 0dBW in an 8Ω circuit is _____ watt/s and corresponds to _____ volts.
3. How many dBm is 100 Watts?

Answers

1. **2 volts**. +6dB is a doubling of voltage and the dBV reference means we compare to 1V.
2. 0dBW is 1 watt regardless of impedance. $P = V^2 / R$. Rearrange: $V^2 = P \times R$. $V^2 = 1W \times 8\Omega = 8$. $8^{0.5}$ = about **2.8V**.
3. **+50dBm**. This one was posed by RF guru Ira Weisenfeld at the 2012 ETA Educators' Forum in Greencastle, IN. You'd be surprised how many well qualified electronics instructors missed it. (I hate to admit that my own first, kneejerk answer was wrong. ☹) It helps to remember that any time you change prefixes, e.g. from milliwatts to watts, you're going up or down 3 orders of magnitude. That's 30dB. So, 0dBW, one watt, is 30dB above 0dBm, one milliwatt, which is to say 1 watt is +30dBm. Then it's two more 10dB jumps to get to 100W. A prefix change between voltage levels works the same way, except that then we're talking about a 60dB increment instead of 30dB; for example 1V is +60dBmV.

Okay, that's almost 9 pages about decibels. I know it was heavy reading, and you should return to it and read it over until dB's are second nature to you. Many techs, even ones with long experience, find dB's a big mystery, so this is an area where you can differentiate yourself from mere mortals.

For the moment, take a break and then we'll look at some other aspects of signals.

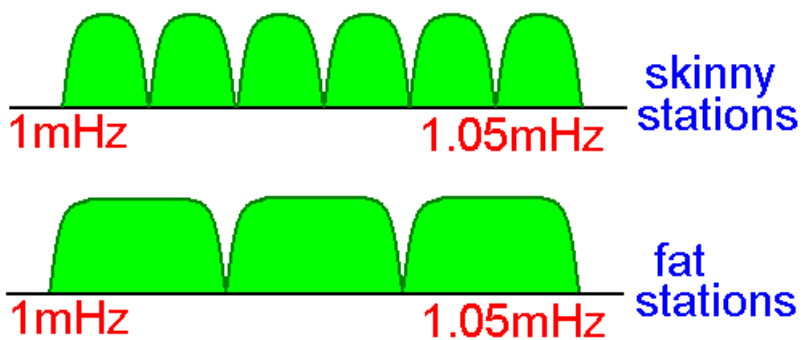
FREQUENCY SPECTRUM

The frequency spectrum of a signal refers to the entire range of frequencies contained in that signal. To quickly recap from EM2, a sine wave contains energy at only one frequency, called the **fundamental**. Every other waveform contains energy at the fundamental frequency and multiples of that frequency, called **harmonics**. Strictly speaking, the fundamental is a harmonic, but the word “harmonics” is informally used to mean everything *but* the fundamental.

Noise and transients, which are normal parts of some kinds of signal, contain energy at every frequency. The concepts of “fundamental” and “harmonics” do not apply to them.

In many cases, it is not necessary to actually preserve and transmit the entire range of frequencies present in the signal. The actual range of frequencies a system must handle, referred to as its **bandwidth**, depends on the type of signal, and also the use that is to be made of it. For example, a hi-def TV picture requires a greater bandwidth than the video link used in a McDonald's drivethru. A system is typically designed to accommodate only as much bandwidth as necessary for a particular application, because the greater the bandwidth, the more expensive the equipment.

Besides the bandwidth limitations of signal generating and processing equipment, every method of moving signal from one place to another – cable, radio, fiber optic, whatever – has its own bandwidth limitations. The more bandwidth each signal occupies, the fewer signals there are room for.



This is one of the fundamental tradeoffs in the design of almost any communication system, analog or digital. It's the reason most AM stations only transmit audio frequencies up to 5kHz. They would sound much better if they transmitted the entire audio band, up to 20kHz – but an AM station transmitting audio up to 20kHz is four times "fatter" than one transmitting to 5kHz, so it would take up four times as much space on the dial. There would be room for only ¼ as many stations as we have now. We'll run into this again in EM6; 2-way radios are using less spectrum these days to make room for more of them on the air.

NOISE AND DISTORTION

As noted in EM2, a signal always experiences some alteration as is passes through electronic equipment. Every stage of amplification, every foot of cable, every connector changes the waveform in some way. Ideally, these changes would be minimized as much as possible. In the real world, that's often a waste of time, effort and money. Instead, a level of performance is chosen that fully meets the needs of the application, and that level is often far lower than the level that's theoretically possible. Thus a cost/benefit analysis is built right into the standard.

For example, a young person can hear frequencies up to nearly 20kHz, and speech contains energy at those upper frequencies. However, everything necessary for clear communication occurs at 3kHz and below. So, the telephone system is designed for about a 3kHz bandwidth. It would be possible to build a 20kHz telephone system, but that would involve a huge price increase and a host of technical problems. It would yield nicer sounding speech, but no better communication.

Since bandwidth, noise and distortion requirements vary by application, instead of discussing them in detail here, I'll discuss their relevance to each type of signal as it comes up, throughout the rest of these study guides.

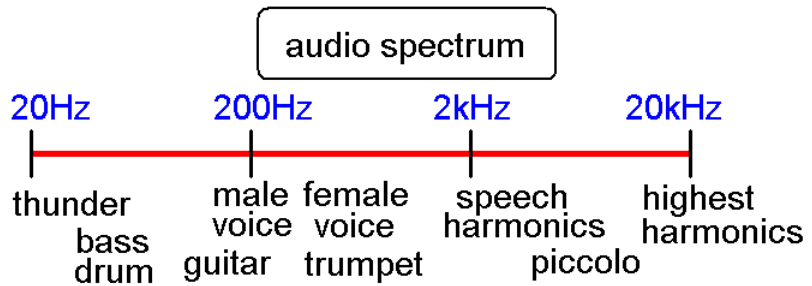
SIGNAL TYPES

We've covered the basic characteristics of signals; now it's time to look at how they map onto specific signal types. Charts which condense this information are included in the printer-friendly Review at the end of this study guide.

AUDIO SIGNALS

“Audio” means sounds, as represented by analog signals, in the frequency band between about 20Hz and 20kHz. These are the approximate hearing limits of a good human ear.

The primary technical problem with audio has always been the 1,000:1 frequency span; it's difficult to build certain things, like speakers, to handle that whole range. Fortunately, there is almost no kind of sound that uses the entire audio spectrum – pipe organ, synthesizers, maybe blockbuster movies – and even in those cases, a more restricted range is usually adequate.



Most audio system interconnections carry **line level** signals. This expression comes down to us from the first audio system, the telephone. It's a practical level for general-purpose signal distribution, neither so high that it's expensive to generate nor so low that it's overly sensitive to hum, hiss and other forms of degradation.

Since a typical audio signal varies in level, it's difficult to pin down what a line level signal is, but in general, a “full” level – a level that would peak at “zero” on an audio level meter – is somewhere between 0.2V and 2V, depending on the specific application.

Telephone

It's hard to imagine a time when the mere transmission of speech over wires was cutting-edge technology, but that was the case in the late 1800s, when the telephone system was being developed. The early telephone system was the result of hundreds of try-this-and-see-what-happens experiments, but in the 1930s, a lot of high-grade science was applied to the problem. The result was a reliable, well-integrated system of voice communication.

I can't think of a more obvious example of a standard that was based on cost/benefit analysis. The telephone system was designed with the absolute minimum performance required to transmit speech clearly. Its characteristics:

- **frequency response**: 300Hz – 3kHz. Although speech has a much wider spectrum, a 300-3k bandwidth is all that is necessary for clear communication. Not coincidentally, that's the approximate bandwidth of an old hand-cranked phonograph.
- **S/N**: often as poor as 20dB.
- **waveform distortion**: 10% or more. Originally, the carbon transmitter microphone was responsible for most of this. Today it's the way the signal is converted to digital and back to analog.
- **signal level**: about one milliwatt, maximum. The telephone system is one of the few places where "line level" is still measured in power terms, because it's one of the few places where a land line directly drives a receiver/speaker. Where the line impedance is 600Ω, the corresponding voltage level is 0.78V. Telephone companies have very strict rules against putting signals much hotter than this on a line, due to the likelihood of interfering with other calls.
- **impedance**: source, line and load impedances are matched. Most often but not always 600 ohms or in that neighborhood.
- **format**: balanced or pseudo-balanced, unshielded, twisted pair. Power to operate each telephone is provided by the line it's connected to. There is a 48V DC potential on the line that is drawn down to about 5V when the handset is picked up.
- **dialing**: originally accomplished by momentarily breaking the circuit which advanced a **stepper relay** at the office, one step per break. Replaced beginning in the late 1960s with the **DTMF** or **dual-tone multi-frequency system**, where tones are used to indicate numbers. These tones are in the same frequency range as speech, but they are very specific, and the telephone system won't recognize a number unless two of these tones are present. That way the dialing system is not accidentally activated by speech.
- **ringing**: a 90V, 20Hz sine wave originally used to mechanically operate a bell.

This basic system is called **POTS** – **plain old telephone service**. Its limitations were already apparent in the 1960s, and by now billions of dollars have been spent trying to shove more and more data through it. We've reached the point where POTS simply isn't up to the task, and most parts of the old system have been replaced by fiber optic and other wideband technologies.

However, sound quality has not improved; in many cases it's gotten worse. The telephone system has always been designed to provide just the bare minimum quality necessary to transmit speech, because the cost of even the slightest improvement, multiplied by hundreds of millions of telephones and a nationwide network, would be huge.

Consumer Audio

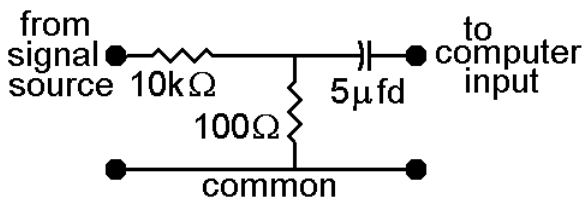
Standards for consumer audio equipment vary a lot, depending on the level of performance required. When people listen to AM radio, they don't expect to hear CD quality. (This is changing in some markets, as some AM stations are "piggybacking" digital audio on top of their regular signal.)

High fidelity refers to a system that produces a sound that's reasonably close to the original – or, in a studio recording built up track by track, where there is no "original," a sound that's close to the sound of the monitor speakers in the studio. (Studio producers will object that this is an oversimplification, and they're right, and please don't shoot me, but this subject is more involved than I want to go into here.) A good "high fidelity" consumer system would have specs something like this:

- **frequency response**: 50Hz to 15kHz. This is limited primarily by the speakers, especially on the low end. Extending flat response even to 40Hz means speakers that are either huge or inefficient (not much sound per watt) and look funny. Just for reference, 50Hz is approximately the lowest "G" on a piano or bass guitar. The "E" below is about 40Hz. If you're old enough to remember the Dolby ad that used to play in theaters, the one with a sliding bass note that goes down to a wall-shaking bottom, that was 40Hz. A speaker specification that shows flat response to 20Hz is almost always a flat-out-fraud: most people have never heard a system that's flat even to 30Hz.
- **noise**: limited by the source. A CD has a dynamic range of over 90dB from floor to clipping level. FM can have a similar dynamic range, but almost never does, because radio stations try to set their controls so the music will be consistently loud.
- **distortion**: as low as possible, thank you. This again is dominated by speaker distortion, which can easily reach 10% at low frequencies even in a good system.

Oddly enough, even when the speaker is cranking out 10% distortion, presence or absence of 0.3% distortion from the amplifier can still be heard. They're different kinds of distortions.

- **level**: the standard line reference level is 0dBV, or one volt, but this is not the actual signal level. Consumer line level is typically around $^1/_3$ volt.
- **impedance**: nonmatched. Outputs are medium-low impedance, inputs are high, so outputs are working into loads that seem like almost-open circuits to them. For transistor and IC equipment, outputs tend to have an impedance of 1kΩ or lower, though if you actually put a 1k load on them, they'll distort. Input impedances are 5kΩ or higher, which is what those outputs are designed to drive. Tube equipment is a law unto itself, which varies from one manufacturer to the next.
- **format**: usually unbalanced, shielded coax cable. Connectors are usually RCA type. If you run into any other kind of connector in consumer gear, check to be sure you aren't dealing with a non-standard signal. Connector exception: computers, and anything computer-related, will use 1/8-inch phone-type audio connectors:
  - Blue connectors: line inputs.
  - Green connectors: headphone outputs. Since headphone voltage is similar to line voltage, these outputs are used as line-level sources too.
  - Pink connectors: microphone inputs. If you feed a line input into one of these, it will likely distort. Also, there's DC on the jack because most microphones nowadays are condenser mics, which require a power source. You shouldn't connect a line-level signal into the pink jack. Unfortunately, you often have no choice: most computers don't have a separate blue input. Here's an attenuator circuit that can be used to adapt a line-level signal to a computer mic input, which I've duplicated in the review pages at the end of this study guide. An attenuator is also called a **pad**.



Just FYI, most built-in computer sound cards are mediocre quality at best. If you're building a studio, use external cards connected via USB.

Professional Audio

Broadcast, recording studio and movie audio technology came directly from the telephone system. Levels are higher and impedances are lower than in consumer equipment because it is more important to avoid signal degradation, because signal degradation is more of a danger (longer cables) and because the added cost doesn't amount to much compared to all the other costs in a studio.

- **frequency response**: again, limited by speakers, but a bit better, usually about 30Hz to 20kHz.
- **noise and distortion**: as low as possible. This is especially critical with pro equipment because, unlike a home system, the signal often passes through a dozen pieces of equipment, each one adding noise and distortion. We'll come back to this in EM4, when we discuss the reasons a lot of pro gear has gone digital.
- **level**: the standard reference level is 0dBu, which is about 0.78V. The actual level most used in pro gear is +4dBu, about 1.23V. The standard level indicator has long been the **VU** or **volume unit** meter, which has become somewhat obsolete but refuses to die. +4dBu reads “0” on a VU meter. Some equipment is calibrated to +8dBu, with the expectation that a 4dB pad will be inserted between the equipment output and the line. A lot of pro equipment has transformer outputs, and the pad is there to give the transformer a predictable load. Not doing so often causes frequency response errors. If you plan to work in pro audio, go back to EM2 and read the material on transformers.
- **impedance**: nonmatched. The actual impedance of a line output is 50Ω or lower, but if you connect it to a 50Ω load you will probably get distortion. A pro output should be able to drive a 600Ω or higher (lighter) load. Most inputs are 5kΩ or higher. An exception is any circuit that takes a signal from or feeds a signal to a telephone line; these must be 600Ω. That's not an issue if the phone company supplies a coupler unit.
- **format**: balanced, twisted pair, shielded. The wires can be unshielded, but because pro audio sound involves much more stringent quality standards than telephone sound, the lines are usually shielded just for insurance.

Low-budget professionals have noticed that consumer equipment sounds pretty good these days, and that has given rise to less expensive **prosumer** equipment which mixes consumer and pro characteristics. For example, a mixing console often has +4dBu balanced main outputs but -10dBV unbalanced inputs and outputs for patching in equalizers etc.

The main problem with mixing pro and consumer equipment is the 12dB difference between -10dBV and +4dBu. (No, that's not a mistake. There's a 2dB difference between 0dBV and 0dBu.)

Just for reference, frequency response of some common audio sources:

- CD: 20Hz low end, very sharply limited to 20kHz at the high end.
- FM: 50Hz low end, sharply limited to 15kHz at the high end.
- AM: theoretically 30Hz low end, though nobody actually goes that low. In most cases, 5kHz high end.
- Downloads: wide variation, from CD quality down to grungy. Depends on how much data compression has been applied to reduce file size.

Both AM and FM stations are capable of "piggybacking" digitally encoded, higher-quality audio on top of their regular analog signals.

With the three exceptions noted below, it is usually safe to assume an audio signal is a line-level signal; that is, on the order of 1 volt.

There are three common types of audio signals which are not at line level:

- **microphone level**, the signal level coming from a microphone, which is much weaker than line level
- **speaker level**, the signal driving a speaker, which is much stronger than line level
- **headphone level**, the signal driving headphones, which is about the same voltage as line level but at a much higher current, resulting in much higher power.

Microphone Level

- **frequency response**: depends on the microphone. Vocal mics are often built to attenuate frequencies below 80Hz, because there are no speech or singing sounds lower than that – only breath, wind and handling noise that we're better off without.
- **noise**: in a well designed system, the self-noise of the microphone is the limiting factor. A 150Ω **dynamic microphone** has a self noise of about 0.25μv. A person speaking 18" away might create a signal of around 1mV, resulting in a 72dB S/N. Hiss would be barely audible if the volume is turned way up. A **condenser mic** can be much better, but usually isn't. There's an amplifier right inside the mic, which establishes the noise floor.
- **distortion**: very low at normal levels. A good dynamic mic can handle an SPL of 140dB, which is high enough to cause immediate, permanent hearing damage. At sensible levels, the mic is loafing.

- **level**: a dynamic mic produces around 1mV from a person speaking 18” away in a normal voice. This is a very rough figure, because sensitivity varies a lot among mic models. When placed inside a kick drum, the same mic may put out 5V peaks. Condenser mics are 3 to 10dB hotter.
- **impedance and format**: in professional work, the standard has long been 150Ω, balanced, shielded. Connection is via 3-pin XLR audio connector. Pins point in the direction of signal flow; thus the male connector is on the microphone end.

  Condenser mics intended for speech reinforcement are often 600Ω. The difference isn't terribly significant because a proper microphone input is not impedance-matched; it should be 3kΩ or higher.

  A mic connected to a belt pack transmitter is usually unbalanced and medium-impedance (a few kΩ) with DC applied to the signal wire at the transmitter end.

  There is also such an animal as a high-impedance mic, around 50kΩ unbalanced, which is dinosaur from the tube era that won't work with modern low-Z inputs. A professional never uses those.

The main difference between microphone level and line level is that microphone level is much lower, unless you're really screaming into the mic. We have no choice about that. Dynamic microphones are headphones in reverse. (The first high-quality headphones actually used dynamic microphone elements.) All the electrical output is generated by actual sound energy. There isn't a lot of power in most sounds, and the microphone only intercepts a small percentage of whatever power there is.

Also, the dynamic range of sounds is very wide. There is no other place in an audio system where the dynamic range is as wide, simply because level controls are used “downstream” to moderate the differences. If a microphone could be built that put out, say, 0.1 volt on soft sounds, it would have to put out 500V on loud ones – quite impractical. So when we're dealing with microphones, we're stuck with low-level signals which have to be boosted before they can be treated as line level signals. An amplifier which boosts a low-level signal to line level is called a **preamplifier**.

Speaker Level

At the other end of the system, significant power is required to drive loudspeakers, much more power than is available from a line-level signal. This is provided by a **power amplifier**, which boosts a line level signal to the point where it can power a speaker.

- **frequency range**: with today's electronics, the range can be as wide as we care to make it. However, very low and very high frequencies are often attenuated to protect speakers.
- **noise**: modern power amps are usually quieter than the signal that drives them, simply because they are placed late in the signal chain where the signal is already fairly hot. So in practice, the noise floor has already been set by the time the signal reaches the power amp.
- **distortion**: the power amp is the main electronic source of distortion in many systems, and whole books have been written which examine the various details. If you work in pro audio, you'll read them. There's still a lot of debate about which kinds of distortions matter and how to measure them.
- **level**: a fraction of a watt to thousands of watts. Depends on how loud the sound needs to be and how efficient the speaker is. A 1-cubic-foot speaker typically requires about a watt, average, with peaks between 10 and 50 watts, to make enough music for an apartment without annoying the neighbors. The biggest mistake people make with power is forgetting to take speaker efficiency into account. A typical bookshelf speaker takes ten times as much power as a paging horn to create a certain intensity of sound. All else being equal, the larger the speaker, the more sound per watt.
- **impedance**: the most common speaker impedance is 8Ω, a very approximate value that varies with frequency. The power amp's output is not 8Ω; usually it's 0.1Ω or lower. In other words, it is a stiff voltage source that's not intended to be impedance-matched.
- **format**: screw or binder terminals designed to connect to a twisted pair, two wires of any kind – lamp cord, flat speaker cable, expensive special cable, whatever. The two things that differentiate speaker cable from other signal cables are, first, it must be reasonably heavy gauge, 18ga. or larger unless we're dealing with really wimpy speakers, short cable runs and low power levels, and second, the wires should *not* be shielded!

Professional systems with a lot of speakers distributed over a wide area – a hospital, for example – usually employ a different signal standard. Instead of being based on a standard speaker impedance like 8 ohms, they're based on a standard amplifier output voltage. The most common standard is 70.7V at full amplifier output.

Headphone Level

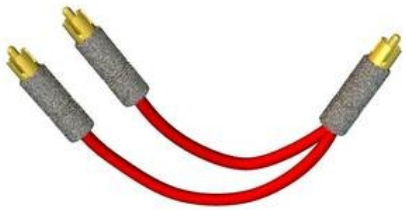
Because headphones are so close to your ears, a tenth of a watt makes a noise that is literally dangerous to your hearing.

- **frequency response, noise, distortion**: like speaker level.
- **level**: 10-100mW. Consumer line level voltage, but much higher current because the load is much lower impedance. Since the voltage is in the same range as consumer line level, a headphone output can be used as a line-level source.
- **impedance**: most headphones are 32Ω. The amplifier output may be lower or higher.
- **format**: the traditional connector is a ¼-inch 3-circuit phone plug, but most headphones today use a $^1/_8$-inch plug. The cord is sometimes shielded, but doesn't need to be; the level is high enough that noise pickup is not a problem.

For iPods and such, the 0.3V output is generated directly. When a power amplifier has a headphone output, it's taken off the much higher-voltage speaker output, but there's about a 100-ohm resistor in series with each channel of the headphone jack. Voltage divider. Yeah, you figured that out before I said it.

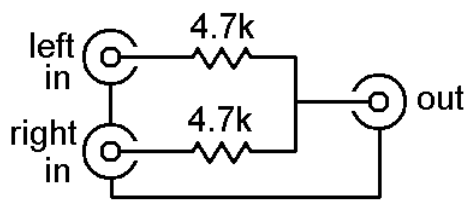
Probably the most important thing to keep in mind with audio signals is the huge variability of levels. Even at times when the level seems constant, there are peaks that far exceed the average. That's why you need 6 to 10dB of headroom, even for non-critical speech applications, and up to 20dB for music.

Thus, if you're running an average level of 0.78V, the equipment should be able to handle 7.8V if you want to avoid clipping.



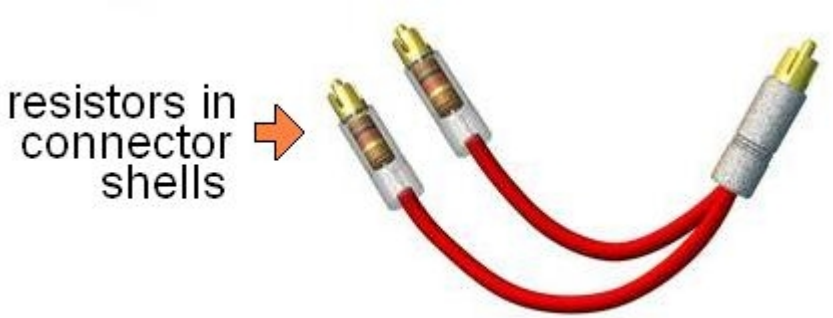
A note about stereo signals: the typical way to combine left and right to feed into a mono input is to use a Y-cord. This is a veddy veddy bad thing! Each output not only contributes its signal to the sum, it also tries to ram its signal backwards into the other output. Result: lots of distortion, sometimes equipment damage.

If you want to sum a stereo signal, you'll have to build your own Y-cord with resistors in it. Why this item isn't readily available off the shelf, when the problem has been known for 50 years, is one of the great mysteries of the Universe. The diagram is for consumer equipment; for pro equipment, use 1k resistors.



combiner cord for consumer stereo equipment

If you assemble the cord with metal-shell connectors, you can put the resistors inside the shells. This will be a good test of your soldering and shrink-tubing skills.



Audio power amp outputs must *never, ever, no way nohow be connected in parallel*. See chapter 5.

RADIO FREQUENCY SIGNALS

- **frequency range, noise, distortion**: depends on the specific type of signal. A car door opener is usually an almost pure single-frequency sine wave. An analog standard-definition video signal occupies 6MHz of spectrum.
- **level**: it is usually impossible to establish standardized levels for RF signals because they so often are transmitted wirelessly, and the received signals vary in strength depending on distance from the transmitter. Even in transmission-line systems, levels vary because cables attenuate high frequencies. RF equipment is designed so that, between specified minimum and maximum limits, level variations either don't matter or can be corrected for.
- **impedance and format**: at high frequencies, even short cables become transmission lines, so RF systems are almost always impedance-matched throughout. RF transmission lines are unbalanced, shielded, 50Ω. A variety of connectors.

At extremely high frequencies, cable losses become so great that it is more efficient to route signals via hollow rectangular tubes called **waveguides**.

There's a lot more to say about RF signals, but I'm not going to say it here, because all of EM6 is devoted to transmitting and receiving signals wirelessly, always involving radio frequencies. However, we're going to run across RF again in later EM3 chapters, as we discuss various kinds of amplifier circuits.

VIDEO

We have to make a distinction between video as it comes out of a camera, called **baseband** video, and video as transmitted over the air, called **modulated** video.

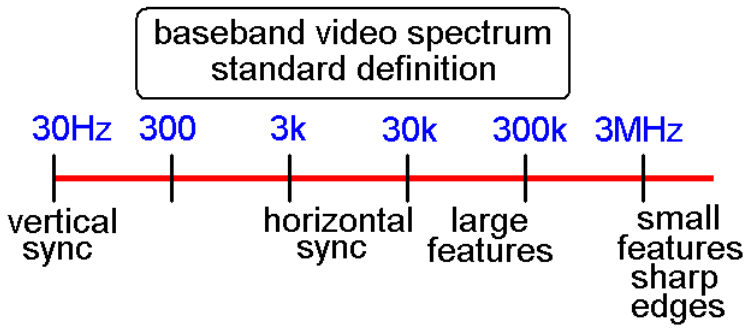
Baseband Video

In an analog video system, inside the camera, the image is repeatedly **scanned** left to right, top to bottom, turning a 2-dimensional picture into a single analog data stream where instantaneous voltage equates to brightness at a particular point. The horizontal scan frequency is much higher than the vertical frequency, resulting in a picture consisting of a lot of almost-horizontal lines. The scanned pattern is called a **raster**.



- **frequency spectrum**: in the traditional standard-definition system, the horizontal scan frequency is just under 16kHz, the vertical frequency is almost exactly 30Hz, and there are 525 horizontal lines. This standard is becoming obsolete in commercial television as broadcasters move to higher-definition formats, but it is still widely used in security and closed-circuit TV.

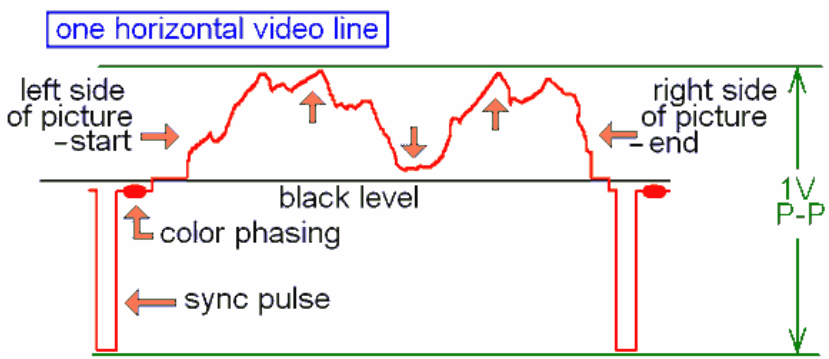
The lowest frequency in a standard-definition video signal is 30Hz. 4MHz is typical for the highest frequency in a standard-definition signal; hi-def can go to 100MHz or higher.



- **noise**: “snow” becomes visible if the S/N ratio is not at least 40dB. Pro studio equipment is usually much better than this.
- **level**: approximately 1V peak to peak.
- **impedance and format**: impedance-matched, 75Ω, unbalanced shielded cable. The most-used connectors are the RCA connector with a yellow insert to identify it as a video connector, and the F-connector, which is a screw-on connector that uses the RG59's signal conductor itself as the signal pin.

Without getting into too much detail, there is a “black level” voltage that forms a baseline for the signal. More positive voltages correspond to various levels of brightness, with a maximum voltage that represents pure white. There are also pulses that go in the negative direction from the base line, used to synchronize the horizontal and vertical scan rates.

Color information is concentrated near the top of the signal spectrum; that is, near 4MHz. Near each negative pulse, there's a short burst of 3.58MHz sinewave. This is used to synchronize the phase of the color decoding circuitry.



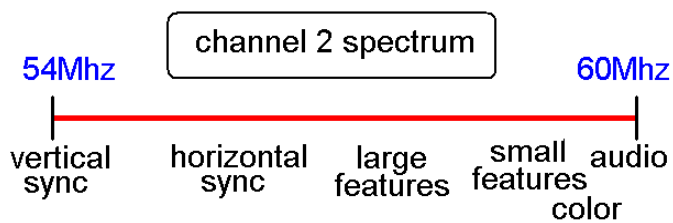
Due to the high-frequency content of a video signal, it must be treated the same way you'd treat RF. If the high-frequency bandwidth of the system is deficient, fine details will be smeared. Color may also be washed out, since the color information is high-frequency, though modern TV's usually have circuitry to correct for this.

Due to the low-frequency content, the 60Hz power line frequency is within the spectrum of a baseband video signal. This means the signal is vulnerable to hum pickup, which shows up as broad horizontal bands in the display. You have to pay attention to ground loops and such.

Modulated Video

When a video signal (or any other signal, for that matter) is sent to a transmitter, it is **modulated** onto a much higher radio frequency which carries it out over the air. This higher frequency is, not surprisingly, called a **carrier**. This analog system is obsolete as far as over-the-air video is concerned, but it is still used in cable systems.

- **frequency spectrum**: a standard-def signal, including audio, occupies a 6MHz chunk of spectrum.
- **noise and distortion**: S/N should be 40dB or better. The most serious distortions are usually timing distortions, which show up as ghosts and blurriness.



- **level**: huge variations, from microvolts to large fractions of a volt. All TV receivers have circuitry to correct these variations.
- **impedance and format**: same as baseband, 75Ω unbalanced shielded. The modulating signal is inverted so the peak transmitted power occurs on the sync pulses and the minimum power corresponds to brightest white on the screen.

Digital Video

Digital broadcast video simply takes baseband video and modulates it differently. Instead of an analog signal where voltage corresponds to brightness, the baseband voltages are coded into digital data. More on that in EM4.

End of EM3 chapter 1/

EM3 CHAPTER 2
ACTIVE DEVICES

Objectives of this chapter:

- *Explain the difference between passive and active devices.*
- *Define: thermionic, solid state*
- *Discuss: semiconductor, intrinsic, doping*
- *Explain how current flows (or doesn't) through a semiconductor.*
- *Define: diode, rectifier, charge carrier, depletion*
- *Discuss a diode's forward and reverse bias behavior*
- *Define and discuss: zener, Schottky, tunnel, Gunn, varactor*
- *Discuss the "valve" concept*
- *Name the parts of a triode and explain how it works*
- *Name the parts of a bipolar junction transistor and explain how it works*
- *name the parts of a junction FET and explain how it works*
- *name the parts of a MOSFET and explain how it works*
- *Define IGBT and discuss its operation*
- *Define "avalanche device"*
- *Discuss the operating behavior of a trigger diode*
- *Discuss the operating behavior of a unijunction transistor*
- *Discuss the operating behavior of a thyristor*
- *Discuss the operating behavior of a triac*
- *Discuss device packages*

This chapter covers enough material that you'll need a road map. The topics we're going to cover, in order:

1. **What an active device is, and how it differs from a passive device**
2. **The two types of active technology, thermionic and solid-state**
3. **Two-element devices, called diodes**
4. **Three-element devices, called triodes and transistors**
5. **Avalanche devices, which act like switches**
6. **Thermistors**
7. **Package types and other physical issues**

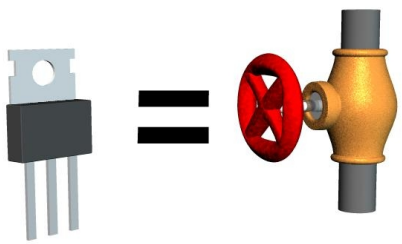
Besides the above, there are also electronic devices that make or detect light, which come under the heading of **optoelectronics**. We'll study those in chapter 10.

As stated in EM1 chapter 1, it's possible to view current as electrons flowing – to +, or positive (conventional) current flowing + to -. I'm going to flip back and forth so you get used to both ways of looking at current flow.

1. ACTIVE AND PASSIVE DEVICES

All the components we've studied to this point have been **passive** devices. They're like mechanical parts: bolts, screws, hinges etc. **Active** devices, on the other hand, are machines, like saws, drills, nail guns etc. They perform operations on electric current – boosting, controlling, modifying, processing it.

The most important thing to know about active devices is that **most simple active devices are current valves** of various kinds. You could say a switch or rheostat is a valve, but those devices are controlled mechanically, so they're electrical, not electronic. A transistor, on the other hand, is a valve that is controlled by an electrical voltage or current – electrons controlling electrons. That is what defines an active device. Electronic valves are used for the same things as mechanical valves, but a thousand other things in addition, and millions of times faster.



2. THERMIONIC AND SOLID STATE

Active devices come primarily in two flavors: **thermionic** devices and **solid state** devices. Thermionic devices depend for their operation on electrons liberated by heat. Solid state devices are crystals of solid material, usually silicon, which electrons flow through.

I have made no attempt to illustrate the actual internal shapes of these devices. Since a tech will never have to take apart a transistor, and has no reason to care what's inside, I think it's more useful to draw these components in whatever way will make it easiest to understand their operating principles.

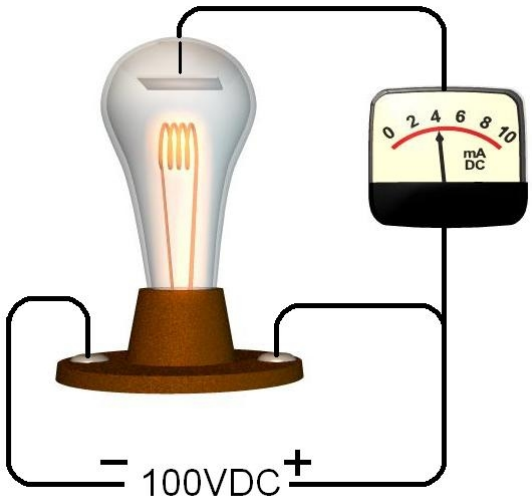
THERMIONIC DEVICES

In a thermionic device, an orange-hot piece of metal emits electrons into a vacuum. We're talking about those old-fashioned, hot glass bottles called vacuum tubes, which were pretty much obsolete by the time your parents hit puberty. I know what you're thinking: isn't it stupid to be talking about tubes in this day and age? Well, first of all, tubes are much easier to understand than many semiconductor devices are. The easiest way to learn about active devices is to study tubes just enough to get the

concepts down, then use those concepts to study solid state devices. Second, tubes refuse to die. High-power transmitters, guitar amps and high-end audio gear still use tubes. Third, thermionic emission turns up in other areas of electronics, so you might as well know how it works.

At temperatures above absolute zero, atoms and molecules are always moving randomly; that's what's responsible for the noise floor we talked about earlier. In a gas, they float free; in a liquid, they stir around like boiling soup; in a solid, they stay in one place but vibrate. If you place a conductor in a vacuum and heat it until it's orange- to white-hot, the molecular vibration becomes so intense that material burns off it into space. That's **thermionic emission**. (therm-EYE-on-ick)

The earliest well-known case of a hot conductor in a vacuum was of course the filament in Edison's incandescent lamp. He noticed that, over time, the inside of the glass bulb developed a black coating that interfered with light output. In an experimental attempt to cure this problem, he ran a third wire to a metal plate inside the bulb. He found that, when he connected a DC potential between the plate and one of the filament connections, with the positive side connected to the plate, a current would flow. He also noticed that the blackening of the bulb stopped. However, when the connection was reversed, so the negative side of the supply was connected to the plate, there was no current.



Edison figured out that the inside of the lamp turned black because material was boiling off the filament. Beyond that, he was at a loss. You can't blame him; at the time, the particles that atoms are made of hadn't been discovered. But today, we know what was going on. The material boiled off the hot filament is primarily electrons and negatively charged ions. When the plate is connected to the positive terminal of the supply, it attracts these freed electrons, accounting for the current flow. It also accounts for the fact that the blackening stopped. And of course, when the supply is reversed, making the plate negative, electrons are no longer attracted to it. The effect doesn't work in reverse because the plate is not hot enough to emit electrons.

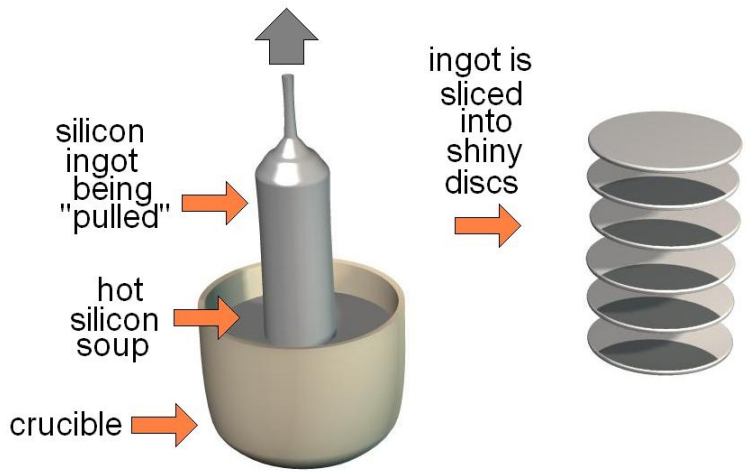
This is called the **Edison Effect**, and it constitutes the first example of a fully electronic active device. Edison didn't realize the importance of what he had discovered, because science wasn't advanced enough. Thermionic emission had to wait until later to be put to use.

SOLID STATE DEVICES

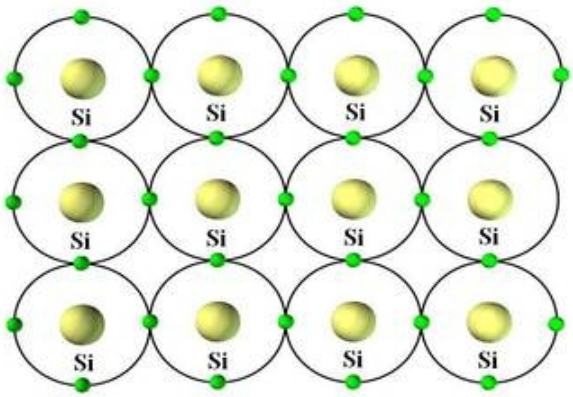
The following discussion is a bit simplified. If I were to fully describe how these devices operate, we'd find ourselves getting bogged down in complications that a technician never deals with anyway. You don't have to understand everything about these devices to get a good idea of how they operate in a circuit. You can find more on the web if you're curious.

Solid state devices are so called because current flow occurs inside solid **semiconductor** materials, instead of in a vacuum. By far the most widely used semiconductor material is **silicon**, the major component of sand and glass. Throughout these EM6 study guides, you may assume "semiconductor" means "silicon" unless a different material is specified.

The manufacture of semiconductor devices begins with a soup of pure, hot, liquid silicon. Material is drawn from this soup which cools and forms a cylindrical ingot several inches in diameter. This ingot is then sliced into shiny discs called **wafers**. A wafer forms a support structure called a **substrate**. Like an artist's canvas, this provides a surface upon which a "picture" is painted. In the case of semiconductors, the picture, produced by photochemical processes, consists of patterns that become transistors and similar devices.



Recall from EM1 that there are conductors, which pass electricity, insulators, which do not, and semiconductors, which have chemical characteristics between the conductors and insulators and whose conductivity depends on circumstances. If it were possible to create a perfect semiconductor crystal with no flaws and no impurities, it would be an insulator. This is called an **intrinsic** semiconductor, a semiconductor material in its natural state.

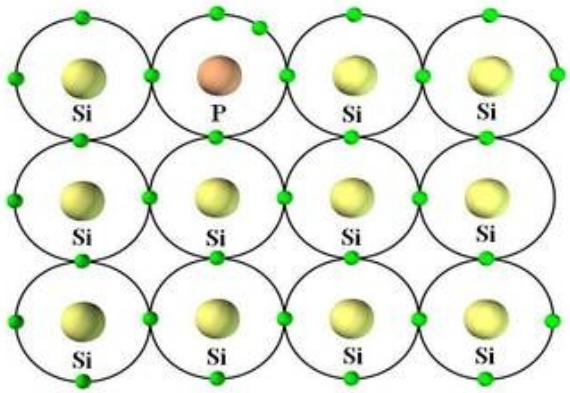


SILICON CRYSTAL

However, a semiconductor with impurities deliberately mixed in is much more useful than an intrinsic semiconductor. Adding impurities to a semiconductor is called **doping.** The impurities are called **dopants** and the resulting materials are called **doped** semiconductors. The percentage of dopants is usually very small, a few atoms of dopant per million atoms of the intrinsic material

.

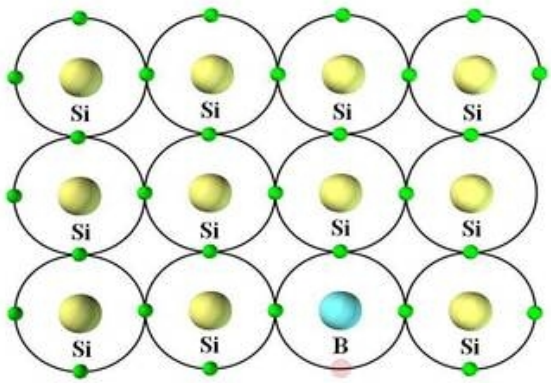
Silicon, the most used semiconductor, has four electrons in its outer shell. If the doping element has a higher atomic number than silicon – for example potassium, which has five outer-shell electrons and a corresponding extra proton in the nucleus – there will be extra electrons that don't fit neatly into the crystal structure as shown above right. A semiconductor with extra electrons is called an **N-type** material.



SILICON DOPED WITH POTASSIUM

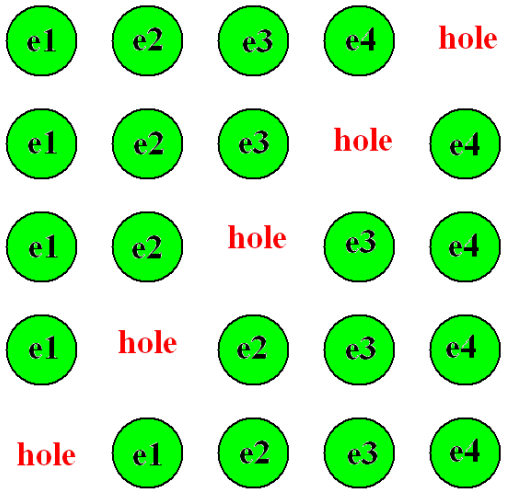
You might wonder why these extra electrons don't just dissipate into the environment, like a static charge. The dopants are whole atoms, with both extra electrons and extra protons. The reason the extra electrons stay around is that there are extra protons to attract them. It is important to understand that the designation “N-type” does not mean that the material is negatively charged. There are an equal number of electrons and protons, so there is no overall charge. But, like a game of musical chairs, some of those electrons have no place to sit. They don't fit into the orderly crystal structure. They belong in the material, but they don't belong in any particular place. Because they are not securely tied to particular atoms, the extra electrons have high mobility – they can move around easily – meaning that a doped semiconductor can be a conductor.

There is an opposite kind of semiconductor material, where the dopant has a lower atomic number than silicon. This means there will be vacancies in the crystal structure where electrons should be, but aren't, as in the boron-doped material below. The reason extra electrons are not pulled in from the environment to fill them is that there is an equal shortage of protons, so the overall charge is neutral. Electron vacancies are called **holes**, and a semiconductor with holes is called a **P-type** material.



SILICON DOPED WITH BORON

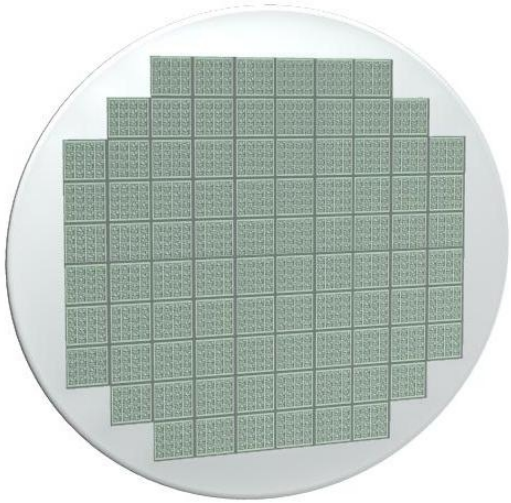
Holes are thought of as positively charged particles that move through the material just like electrons. This is not entirely imaginary. As shown in the illustration above right, when electron e4 moves to the right to fill a hole, it leaves another hole behind. When electron e3 moves to fill that hole, a new hole is created again. Since one hole is the same as another, this is the same as one hole going to the left, moving through the semiconductor like a bubble.



Holes and extra electrons in a semiconductor material are called **charge carriers** – charged particles (or, in the case of holes, virtual particles) that move through the material, carrying their charges with them. If the semiconductor is connected across a voltage, electrons will move toward the positive end and holes will move toward the negative end. Though charge carriers only account for a few parts per million in the silicon, they're enough, in many cases, to move hundreds of amps of current.

Let's go back to those silicon wafers. If we deposit thin, patterned layers of P- and N-type silicon on these wafers, several layers deep, we get a bunch of little semiconductor sandwiches. Where P- and N- layers meet, they form **PN junctions**.

A single substrate wafer typically contains thousands to millions of these devices. After manufacture, the wafer is cut apart and the individual devices placed in casings with terminals to connect to the outside world.



PROCESSED WAFER BEFORE CUTTING

A semiconductor junction conducts current under some conditions and blocks current under others. It is the foundation upon which the modern electronic industry is built.

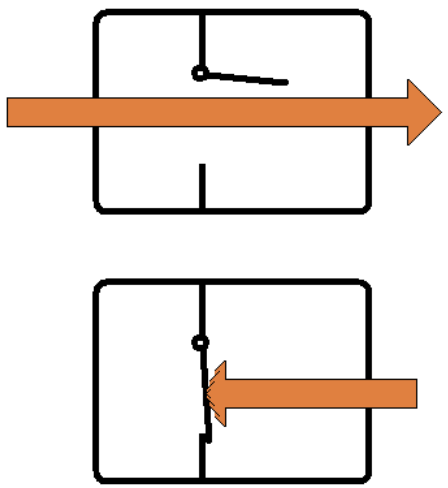
TYPES OF ACTIVE DEVICES

All active devices use the same thermionic or solid state technology. Their various functions are derived from the number and kinds of thermionic elements in the vacuum, or the number and kinds of semiconductor junctions, and how they are arranged.

3. DIODES

The simplest active device is the **diode**. Many people would argue that, strictly speaking, it is not an active device because it doesn't have a control terminal. I prefer to lump it with active devices because a) it performs a control operation on voltage and current and b) it shares the structure and technology of other active devices.

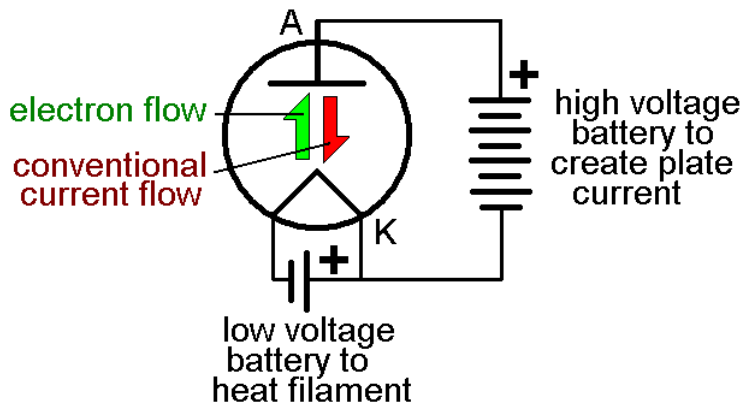
A diode is an electrical one-way valve; current can only flow through it one direction, like the turnstile at the entrance to a stadium. Diodes used in high-power circuits are more commonly called **rectifiers**.

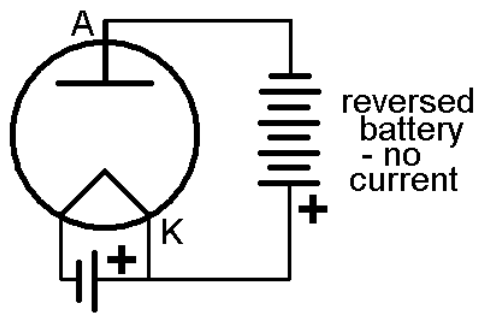


"Di" is a prefix meaning "two;" a diode has two parts. The two active parts of a diode are the **anode** and **cathode**. The diode permits conventional current to flow in the anode-to-cathode direction, but blocks it in the cathode-to-anode direction.

THERMIONIC DIODES

The earliest thermionic diode was the **Fleming Valve**, which is nothing more than an Edison Effect lamp put to good use. As explained above, the conditions for electron flow in a vacuum are 1) an orange-hot **filament** to emit electrons, and 2) a conducting **plate** nearby with a positive voltage to attract them, as in the upper illustration below. If you put a negative voltage on the plate, as in the lower illustration,

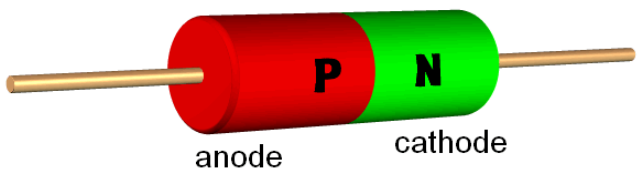




current will not flow because the plate is not red hot and thus does not emit electrons. This means we have a thermionic one-way valve. If we were to compare electricity to water flowing through pipes, a diode would be a check valve.

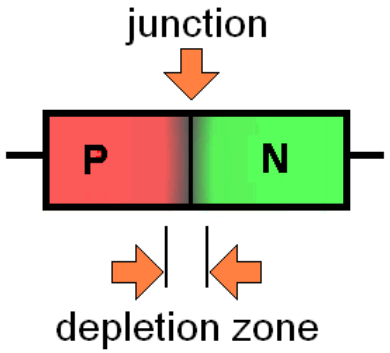
The plate is called the **anode** and the filament is the **cathode** (we use K because C is already used for Capacitance). Electrons flow from filament to plate, which is to say that conventional current flows from plate to filament, anode to cathode.

SOLID STATE DIODES



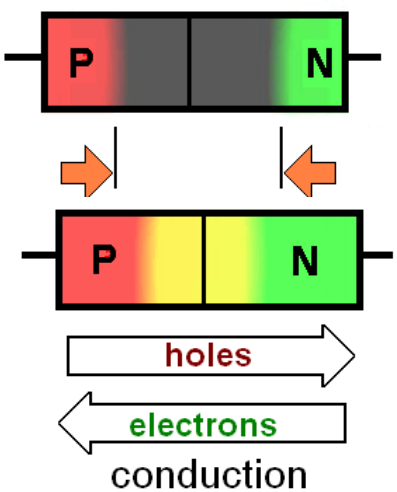
A solid-state diode is made by joining a P-type semiconductor with an N-type semiconductor. The P-type region is the anode, and the N-type region is the cathode. To understand the operation of a solid state diode, let's look first at a PN junction with no external voltage across it.

In most of the material, the extra electrons and holes stay where they are. However, near the junction, they are close enough to attract each other. The electrons fill the holes, resulting in a zone without free charge carriers. This is called the **depletion zone**, a zone depleted of charge carriers.



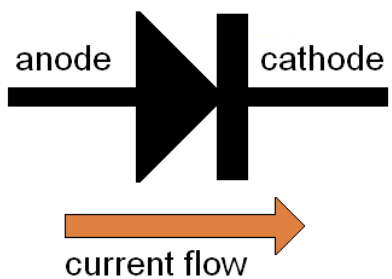
Suppose a voltage is connected across the diode, so the negative voltage is on the anode. The negative voltage will attract the holes in the P-material away from the junction and the positive voltage will attract the electrons in the N-material, creating a wider depletion zone. Since there are no carriers in this zone, current will not flow.

Now let's reverse the connections, and connect the positive voltage to the anode. The positive voltage will repel the holes in the P-type material, driving them toward the junction. The negative voltage will repel the electrons in the N-type material, driving them toward the junction. The electrons and holes sort of "pass through" each other, like ghosts, combining and then separating on their way to the positive and negative terminals respectively; thus current flows.

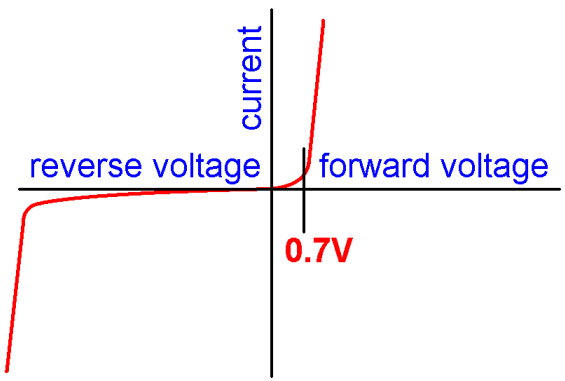


When a circuit is arranged so the positive end of the supply is on the anode side of the diode, so current can flow through it, the junction is said to be **forward biased**. When the positive end of the supply is on the cathode side, so no current flows, the junction is said to be **reverse biased**.

The schematic symbol for a diode combines an arrow and a bar, with the arrow pointing in the direction of conventional current flow. The anode is the arrow end of the symbol and the cathode is the bar end. One of the reasons so many people find it easier to think in terms of conventional current is that diode and transistor symbols were invented by engineers who though in terms of conventional current, and the arrows point in the direction of conventional current flow.

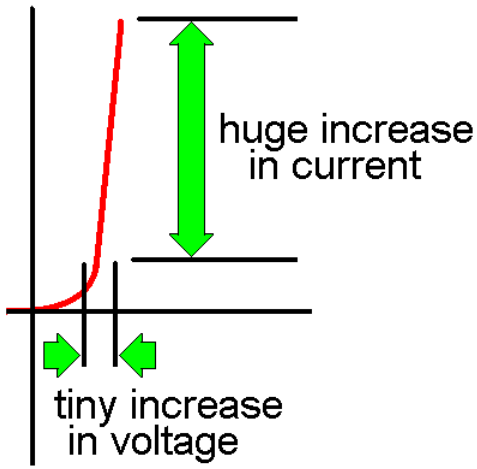


Let's look at the voltage-vs-current characteristics in more detail, starting with forward-voltage behavior. As you can see, it takes a small amount of voltage in the forward direction to close the depletion zone and make current flow.



Think of a one-way door with a spring holding it closed; it takes a small push for it to open, and you have to keep pushing to keep it open. The electrical equivalent of push is of course voltage. It takes a small forward voltage across the diode to make current flow, and regardless of current, that small voltage drop is always there. This voltage depends on what the diode is made of. For silicon, it's about 0.6 to 0.7 volts. We can think of a **forward-biased** diode as a rather good conductor with an unchanging 0.7V drop between anode and cathode, like a broad sidewalk with a step in it. (Some solid-state devices are made of germanium, which has a forward potential of 0.2 to 0.3V.)

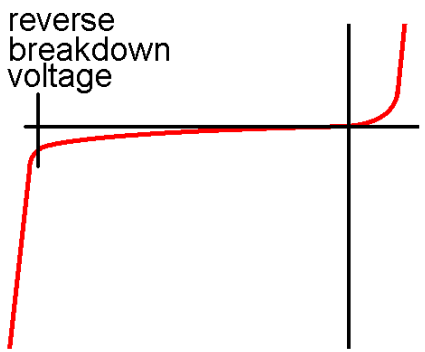
Once the threshold is exceeded, a very small increase in voltage will result in a huge increase in current. You simply cannot make a diode forward-drop much more than 0.7V – you might say it "shorts out" any increase.



A diode has no built-in current limiting function. If you try to put too much current through it, it will try to pass it, until it melts. It is up to the user to look at the manufacturer's specifications to insure that a particular diode can handle the current required in a particular application. Fortunately, diodes – large ones especially – can take short-duration hits far in excess of their continuous current capability.

Now let's look at the **reverse biased** behavior. In an ideal diode, when a reverse voltage is applied across the diode, no current would flow. However, in a real diode there's always a small amount of "leakage" which increases with voltage. This leakage may or may not matter depending on the application. If it does, there are low-leakage diodes available.

If we continue to raise the reverse voltage, we'll reach a point where the diode starts to conduct. This is called the **reverse breakdown** voltage. In contrast to the situation where an insulator breaks down, becoming something close to a short circuit, a diode simply begins conducting, passing as much current as necessary to pull the voltage down to the reverse breakdown voltage. If the applied voltage falls below this value, the diode will stop conducting. In effect, the diode prevents the voltage from rising further by "spilling" current, just as a glass will spill liquid if you try to fill it higher than its brim.



If we're dealing with any significant amount of power, this is likely to destroy the diode. Remember, power is voltage times current – in this case, the voltage across the diode times the current through it.

Let's take an example. Suppose we have a small plastic-cased diode, which can dissipate about a watt, and suppose it has a reverse breakdown of 50V. Suppose we pass ½ amp through it in the forward direction. Power dissipated by the diode is:

$$0.7V \times 0.5A = 0.35W$$

so there's no problem. Now suppose we pass ½ amp in the reverse direction. Power dissipation is:

$$50V \times 0.5A = 25W$$

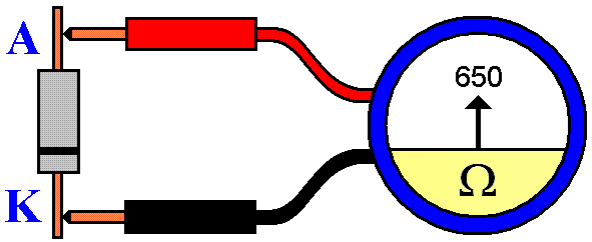
Meltdown!

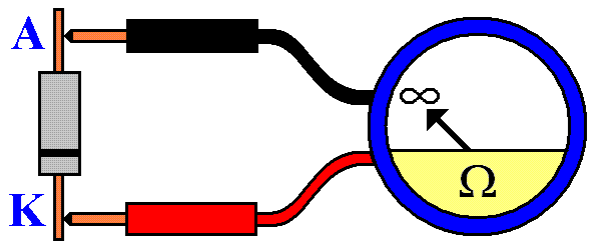
Obviously, when selecting a diode, you have to be sure to choose one with a reverse breakdown voltage that's higher than any reverse voltage that will ever appear at that point in the circuit. Keep in mind that an AC sinewave peaks at 1.4X its RMS voltage.

TESTING A DIODE

A diode can be tested with an ohmmeter. If you've gone thru the EM1 study guide, you know that an ohmmeter tests resistance by putting a voltage across its test leads and looking for current flow. In most ohmmeters, the voltage is 1.5V and the red test lead is positive. 1.5V is high enough to make a diode conduct, but of course only in one direction.

Set the meter to a medium range, e.g. 2kΩ. If you put the red lead on the cathode (in a glass or plastic diode, the end with the band around it) you should read a very high resistance, almost an open circuit. When you reverse the leads, putting the black lead on the cathode, you should get a medium resistance, showing that the diode is conducting – but *not* zero resistance. Remember, there's a 0.7V forward drop across the diode; the meter will interpret this as resistance. Typically, using the 2k scale, the meter will read between 500 and 1k ohms. Many meters have a specific diode-measuring setting.





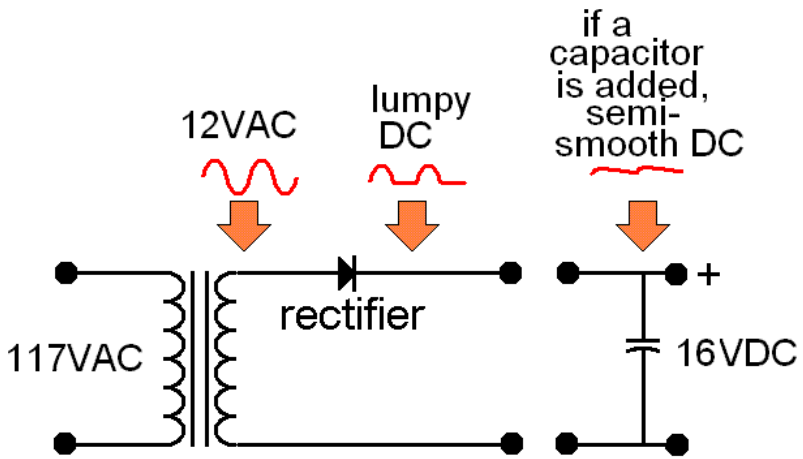
A good diode has a high **front-to-back ratio** – that is, a very large difference between forward current and reverse current. Some people talk in terms of resistances, but that's like using a thermometer to measure weight. With a diode, we're not really measuring resistance but voltage drop. The important thing is that a diode should conduct in the forward direction a thousand times better than it conducts in the reverse direction. That's a good diode. If you don't know which is which, with a particular diode, this test will sort that out for you.

A COMMON DIODE APPLICATION

In most electronic equipment that plugs into a 117VAC outlet, the circuitry actually operates on DC at some different voltage. This means the unit must include a circuit which converts 117VAC to some other voltage, turns it into DC and smooths it out. That kind of circuit is called a **power supply**, shown on the next page.

We'll start with a transformer to convert 117VAC to something close to the voltage required by the load circuitry. Since we're dealing with significant power, we will follow common practice and refer to our diode as a **rectifier**, which is what high-current diodes are typically called, and we'll say the rectifier **rectifies** the AC. In a power supply, a rectifier blocks the AC current in one direction, resulting in lumpy DC. Then a capacitor is added to smooth out the DC.

At the waveform peaks, current flows through the rectifier, charging up the capacitor. Then, the stored charge is drawn off by the load until $^{1}/_{60}{}^{th}$ of a second later, when the next peak comes along and charges up the capacitor again.



Note that the capacitor charges up to the peak voltage of the waveform, minus the 0.7V drop across the rectifier. From chapter 2, we know that a sine wave's peak is about 1.4 times its RMS value. This means the DC output voltage from this kind of supply is *higher* than the RMS value of the AC voltage it comes from. In this case, the transformer's AC output is 12V but the final DC output is about 16V. This power supply is a simple design that would be adequate in many real-world situations.

SCHOTTKY RECTIFIERS



When the voltage across a diode is near zero, it has a significant amount of capacitance. Because of this, it takes the diode a significant time to turn on when forward biased, and even longer to turn off when reverse biased, especially when the diode is brawny enough to be called a rectifier. (Turning on and off is called **commutation**.) Think of a valve that takes a while to open and close because it's full of axle grease. The bigger the rectifier, the harder it is to make it turn on and off quickly. This isn't a big problem at 60Hz, but at higher frequencies it's a real headache.

A **Schottky rectifier** is a kind of “deluxe” rectifier with much lower capacitance, thus faster switching times. Even big, high-power Schottkys can switch quickly; usually at least an order of magnitude faster than regular rectifiers of the same current rating. Also, a Schottky diode has a forward voltage drop much lower than 0.7 volt.

Together, these characteristics reduce the amount of power wasted in the AC-DC conversion process. In chapter 9, when we study power supplies in detail, the Schottky's advantages at high frequencies and low voltages will become evident. The symbol is a diode with hooks on the ends of the bar, suggesting an “S.”

SUPERBARRIER RECTIFIERS

Schottky rectifiers have higher leakage current when reverse-biased than regular rectifiers. The leakage current is still so much smaller than typical current when forward biased that it can be ignored in most applications. Where reverse leakage is not acceptable, **superbarrier** rectifiers are used. As the name implies, they do a super job of rectifying. They are essentially Schottky rectifiers, but with the lower reverse leakage of a regular silicon diode.

The basic rectification function we've been discussing is pretty simple. However, diodes have hidden talents. Engineers have found clever ways to take advantage of their limitations, turning faults into virtues, making them serve other functions you'd never have expected.

ZENER DIODES

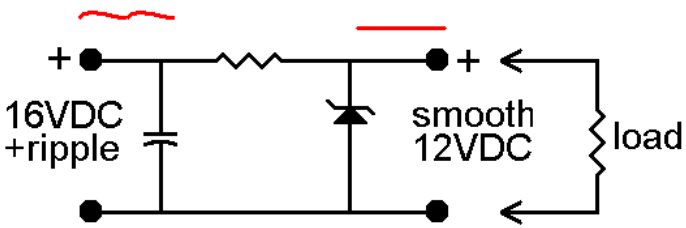
In an AC-to-DC conversion circuit like the power supply on the previous page, the DC voltage varies with variations in the AC power line voltage and with loading, and contains a small residual AC voltage called **ripple** that results from imperfect filtering. This messy DC can be cleaned up and regulated by taking advantage of a diode's reverse breakdown characteristic.

Recall that there's a limit to a diode's ability to block reverse voltage, and when that voltage is exceeded, the diode starts conducting. It will spill as much current as needed, dying a hero's death if necessary, to keep voltage from rising higher.

A zener diode is simply a regular diode whose reverse-breakdown characteristic has been optimized for use as a voltage regulator. The reverse voltage vs. current curve is very flat and the knee where it begins to conduct is very sharp. In the schematic symbol, the anode bar has bent ends. It's pronounced “ZEEner.”



A zener regulating circuit consists of a series resistor and a shunt zener. In other words, it's a voltage divider, except that instead of the output voltage being a percentage of the input, it's a fixed voltage equal to the zener's reverse breakdown voltage.



Note that the zener is in the circuit backwards, pointing in the direction opposite to conventional current flow. That's because we're making current flow through the diode in the reverse direction and using its breakdown voltage to regulate.

The zener shunts as much current as necessary to hold the output to a specific unvarying, unrippling voltage. If it were put it into the circuit the normal way, with current flowing in the forward direction, it would pull the output voltage down to 0.7 volt like any other diode.

The only requirements are that we start with a DC voltage that's higher than the zener voltage, and we choose an appropriate series resistor. Its resistance must be low enough to pass a little more current than the load will ever need, so there's always some current flowing through the zener – but not enough current to burn the resistor or zener out. This involves straightforward Ohm's Law and Watt's Law calculations.

Let's find the resistor value for the circuit above. Suppose the peak ripple on the DC is + and – 1V; that is, the input voltage varies between 15V and 17V. We want a clean 12V out, so we've chosen a 12V zener. At the input voltage's lowest point, the resistor must drop the 3 volts between 15V and 12V. Now suppose the load draws 10mA. Let's set the current through the resistor at 15mA when the input voltage is 15V, insuring that at least 5mA will flow through the zener at all times.

$$\mathbf{3V / 15mA = 200\Omega}$$

We will take this as the maximum acceptable resistance; anything higher might let the output voltage drop below 12V, which would let ripple through.

Now let's calculate the minimum acceptable resistance. If we use a common 1/2-watt zener, the maximum current it can handle without burning out is:

$$\mathbf{0.5W = VI = 12 \times I_{MAX}}$$
$$\mathbf{I_{MAX} = 0.5 / 12 = 0.083A = 83mA}$$

If the average DC input is 16V, the resistor will have to drop an average of 4V. If we let the resistor pass a maximum of 83mA to the zener and 10mA to the load, its value will be:

$$\mathbf{83mA + 10mA = 93mA}$$
$$\mathbf{4V / 93mA = 43\Omega}$$

Thus: any resistor between 43Ω and 200Ω will do the job. Lower than 43Ω, the zener will melt. Higher than 200Ω, and there's a chance the output voltage will momentarily fall below 12V when the input ripple waveform is at its lowest point, 60 times per second.

Naturally, we don't want to waste power or stress the zener, so we'll choose a resistor that's closer to 200Ω. A 150Ω resistor will give us added safety margin against an unexpectedly low input voltage.

Zener regulation is an inherently wasteful process. The power supply provides more current than is needed, and the zener flushes a portion of it into the ground rail. This isn't as serious a problem as it seems. The parts of a circuit that require the cleanest DC power are the low-powered parts that handle low-level signals. Wasted power is only a few milliwatts. High-powered parts of the circuit usually handle such large signals that a little ripple on the power supply doesn't matter, and those circuits can often be designed so they are “blind” to ripple; thus they may not need regulation.

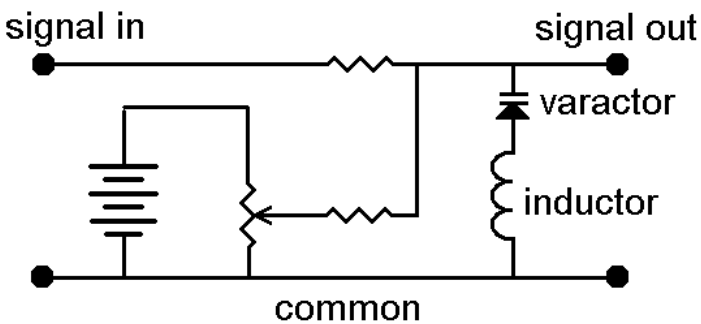
VARACTOR DIODES

Suppose you want to adjust a variable capacitor, but instead of twisting the shaft by hand, you want to control the capacitance electronically, by a computer for example. You could attach a motor and turn the capacitor's shaft by applying a voltage to the motor, but that's pretty clumsy. It would be much better if capacitance could be adjusted directly, via electronic control.

A diode's capacitance varies with voltage. It's greatest at small forward-bias voltages, around 0.4V, just below the diode's conduction threshold. If we reduce this voltage to zero, and then keep going into the reverse-biased region, the capacitance will get smaller and smaller. It will be at its minimum when the reverse voltage is maximum, just before breakdown. If we put the diode in a circuit where it's reverse biased, with a significant DC voltage across it, we can electronically change the voltage to change the capacitance – no motors involved. Any diode can be used for this, but specially made **varactor** diodes work best.



The symbol is a diode incorporating a capacitor. This doesn't mean there's an actual capacitor in series, or that the diode won't conduct in the forward direction – it will, just like any other diode. But it's not intended to be used that way. It's intended to be used with a reverse DC voltage across it, combined with a much smaller AC signal voltage.

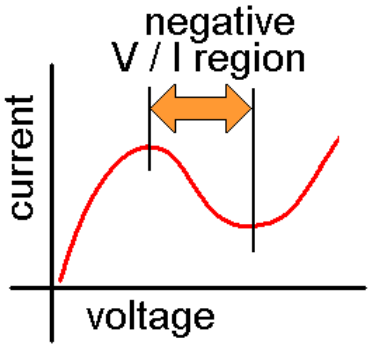


The varactor and inductor form a resonant circuit that puts a notch in the frequency response. The DC voltage across the varactor, adjusted by the pot, controls the varactor's capacitance and thus the notch frequency.

TUNNEL DIODES



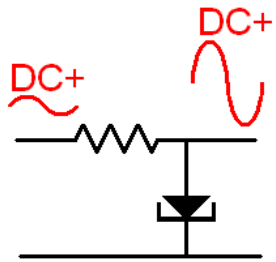
The word “tunnel” refers to a special characteristic of the diode's construction, which we're not going to go into because it doesn't concern us. What does concern us is the diode's voltage vs. current curve. There's a part of the curve where, if the voltage is increased, current *decreases*. In other words, the diode has negative resistance.



If the voltage across the diode is supplied through a resistor, then when the voltage is increased, current will decrease, leading to a further increase in voltage. Thus any change in voltage will be magnified, leading to instability. When we're designing a circuit to generate a continuing waveform, instability is a good thing. We want a circuit which, like a buzzer, refuses to find a stable resting state.

Circuits that generate continuing waveforms are called **oscillators** and we'll be studying them in chapter 6. Tunnel diodes have been used since the 1960s to build oscillators that produce extremely high frequencies, e.g. 10GHz. They are less often used now that we've learned how to make transistors that work at those frequencies.

Tunnel diodes can also amplify. The circuit at right looks like an attenuator, but if the shunt element has negative resistance, any change in the input will result in a greater change at the output – in other words, gain.



Gunn diodes are another type of diode with a negative resistance characteristic, used for the same purpose.

DIODE OPERATION REVIEW

- A diode is a 2-terminal device consisting of an anode and a cathode. A diode built to handle current of 1A or higher is more commonly called a “rectifier.”
- A diode is a 1-way valve. When the anode is positive with respect to the cathode, it conducts, in which case it is “forward biased”; in the opposite case it does not, in which case it is “reverse biased.”
- When forward biased, there is a somewhat-constant voltage drop across the diode. The forward drop for a standard silicon diode is about 0.7V. When the forward voltage is lower than this, the diode does not conduct.

- There is a limit to the reverse voltage a diode can block. When the voltage begins to exceed this limit, conduction occurs. This is called “reverse breakdown.”
- A diode has capacitance, especially at low reverse voltages. High-current diodes have the most capacitance, which makes it difficult for them to rapidly start and stop conducting as the polarity of the voltage changes.
- Schottky diodes have a lower forward voltage drop than the usual 0.7V and reduced capacitance, thus they are better in high-power, high-frequency switching applications.
- Super-barrier diodes have the Schottky's speed and low forward drop, with lower reverse leakage.
- A zener diode is a regular diode whose reverse breakdown characteristic is optimized for voltage regulation applications.
- A varactor is a regular diode whose capacitance vs. reverse voltage characteristic is optimized for electronically variable capacitance applications.
- Tunnel and Gunn diodes have a negative region in their voltage vs. current curves, making them unstable. This instability is used to create high frequency oscillators.

DIODE HANDLING AND USAGE PRECAUTIONS

- Diodes are not static sensitive.
- A glass diode can easily be cracked by bending the leads against the case. The proper method is to bend them over something else; the edge of a printed circuit board often gives the right spacing.
- The published forward current should not be exceeded for any significant time. However, power rectifiers can handle a forward current many times greater than their continuous current rating for a few milliseconds. Manufacturers publish these figures.
- The reverse breakdown voltage should never be exceeded. Doing so will not necessarily destroy the diode if current is limited to a safe value, but if you're deliberately using the reverse breakdown to regulate, you should be using a Zener.
- In applications where the combination of current and forward voltage drop would result in diode dissipation greater than about a watt, heat sinking must be provided.
- Solder quickly and efficiently. Molten solder is hot enough to melt silicon. The way to keep that from happening is to touch the iron to the leads for only a few seconds. When you get to the EM5 study guide, you'll find a whole section on soldering.

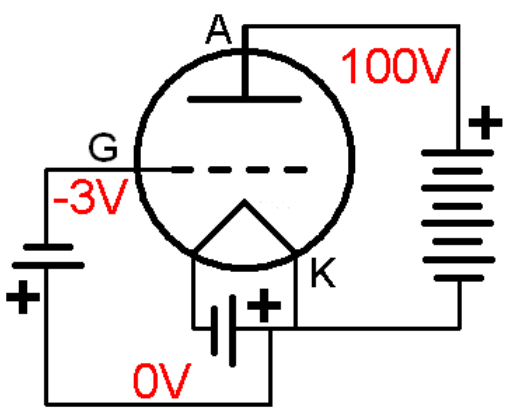
4. TRIODES AND TRANSISTORS

A diode is a one-way valve that just does what it does; it opens when current flows in one direction and closed when it tries to flow in the other direction. That's very useful, but it would be even more useful if the diode had a "control handle," a third connection point, so its opening and closing would be under our control.

If we take the 2-element diode and add a third element, it becomes a **triode**. The added element can be used to turn current flow on, partway on, or off.

THERMIONIC TRIODES

We've seen that, if we have a hot filament and a plate in a vacuum, and we put a voltage between the plate and one side of the filament, with the plate positive, current will flow. We can't control the current; it will flow as long as there's a hot plate and a positive potential. In 1906, Lee De Forest discovered that an open wire mesh or grid inserted between the filament and the plate can be used to control the current. If the grid is left at zero volts, there will be little change in the plate current. But if the grid is made negative with respect to the filament, current will be reduced. And if it is made negative enough, current will stop.

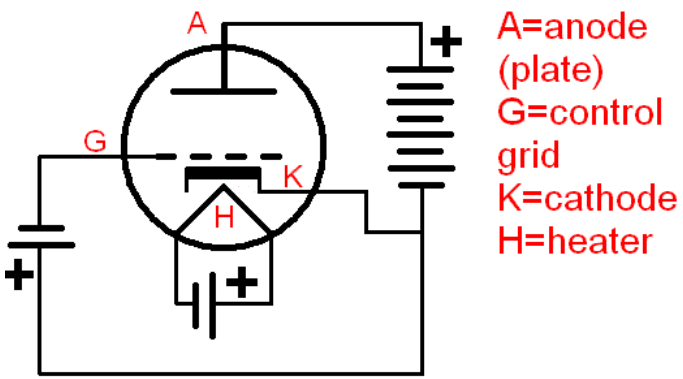


It's not hard to figure out how this works. The grid ("G" in the diagram at left) is mostly open space, so if it is at zero volts, most of the electrons will pass right through it and go on to the plate (anode). But if it's made negative with respect to the cathode, it will create a negative field that repels electrons.

Since it's between the filament and the plate, a fairly small negative voltage will be enough to override a large plate voltage and suppress electron flow. In actual tubes, the grid usually takes the form of a mostly-open coil of wire surrounding the cathode.

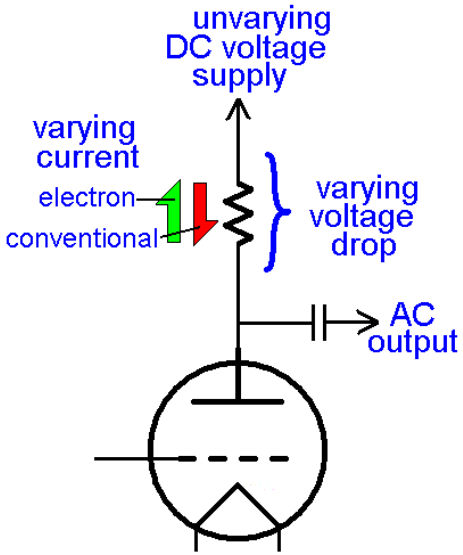
The triode lets us use a small change in grid voltage to cause a large change in plate current. In other words, a voltage-controlled **valve**, which is what the British have always called the triode tube.

The three parts of a triode are the **cathode** (the filament), the control **grid**, and the **anode** (the plate). In more advanced tubes, the cathode is a separate part, indirectly heated by the filament but not electrically connected to it. The filament, now called a **heater**, provides the necessary heat but is not part of the valve circuit.



Normally, we prepare a triode for use by supplying it with a small negative voltage on the grid, which allows a small plate current to flow. This is called **setting the operating point** or **biasing** a tube. Then, if we superimpose an AC waveform on the DC voltage at the grid, so it varies a little above and below the operating point, we will get an AC current waveform superimposed on the plate current.

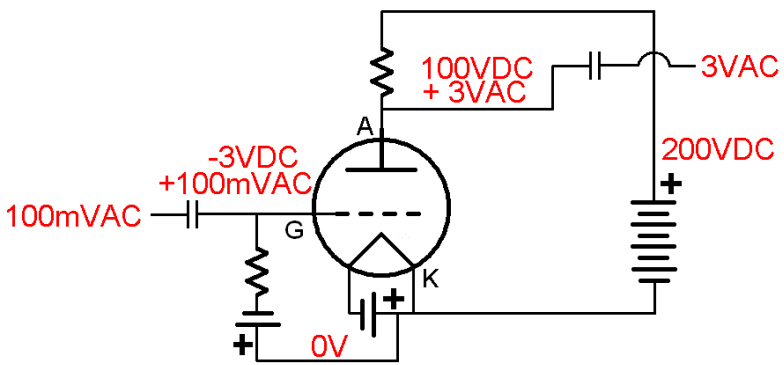
Note that the grid voltage controls plate *current* – not plate voltage. That means, if we apply an AC signal voltage to the grid, we get an AC current variation at the plate. Probably not usually what we want, as signals are usually measured in volts (or watts, if we're making a power amplifier). So we have to convert the varying plate current to a varying voltage. We do this by simply inserting a resistor between the plate and the power supply. As plate current varies, so does voltage drop across the resistor. This gives us an AC signal output.



TRIODE AMPLIFIERS

If everything is done correctly, the output will be larger than the input. Thus we have a device which, until the invention of the triode, was desperately wished for but as imaginary as flying pigs: an electronic device that makes signals stronger.

Terms: the strength of a signal is called **amplitude**. The process of making a signal stronger is called **amplification**. A device that amplifies is called an **amplifier**. The amount of amplification is called **gain** and it can be measured in multiples (an **amplification factor** of, say, 100X) or in dB. An amplifier can be designed to amplify voltage, current or both. Since power is voltage times current, an amplifier that boosts either voltage or current also boosts power.



In the example shown above, the grid is at an average voltage of -3VDC, with a 100mV variation that represents the input signal. The plate is at an average voltage of +100VDC, with a 3V variation that represents the output signal. That's a signal gain of 30X or roughly +30dB.

(The opposite of amplification or gain is **attenuation** or **loss**.)

A given change in grid voltage will result in a given change in plate current. That's called **transconductance**. The symbol is g_m. The measurement unit used to be the **mho** – conductance being the reverse or complement of resistance – but now that unit has been given the name **Siemens**, abbreviated **S**, which you could confuse with seconds if you're not paying attention. All else being equal, the greater the transconductance, the greater the voltage gain. Even if you think you'll never deal with tubes, you should memorize those names, because there's also a kind of transistor where gain is measured in Siemens.

Note that, under normal conditions, current flow in the grid is zero. After all, it's negative with respect to the cathode, so it repels electrons. There is some capacitance between the grid and the other elements, but otherwise it's an open circuit.

A triode requires a fairly high voltage, like 100V, to operate. When the anode-to-cathode voltage drops below a certain point, typically 50V, the anode no longer has enough “pull” to attract electrons.

I think future historians will view the 1906 invention of the triode as one of the great turning points in history. To engineers, having the triode was like being freed from a straitjacket. In the early days, the most aggravating problem radio engineers faced was that there wasn't any way to make a weak signal stronger.

If you wanted to listen to a radio station, the only signal you had to work with was whatever tiny part of the actual transmitted power you could pull out of the air. You strung up a long outdoor wire antenna, put on headphones, and hoped you could pick up enough signal to hear. If not, there wasn't anything you could do. And at the station itself, the impossibility of amplifying audio signals meant, for several decades, that the only kind of signal a station could transmit was Morse code.

But with a triode, if a signal wasn't strong enough, you could make it stronger. If it still wasn't strong enough, you could send it through another triode, and another, almost without limit. The ability to boost a weak signal meant people could receive radio transmissions at great distances. It also meant engineers could stop worrying primarily about having enough signal, and turn their attention to other things like signal quality, flexible control and so on.

But that's ancient history, right? Not really. Transistors are valves too. Though they are built differently and operate differently from tubes, and have shrunk to microscopic size, the same principle – electronic control of current flow through the device – is the foundation of all modern electronics.

TRIODE OPERATION SUMMARY

The basic operation of a triode is really simple:

- A triode is a current valve.
- Current flow is voltage-controlled.
- The triode's components are the cathode, which supplies electrons, the anode or plate, which collects them, and the grid, which controls current flow.
- Conventional current flow is from anode (plate) to cathode.
- A triode is normally “on.” To turn it off, a negative voltage is applied to the grid.
- A triode usually requires 100V or more at the plate to operate.

Some tubes have four or five elements, which makes them **tetrodes** and **pentodes**. They operate like triodes, but with higher gain and other improvements.

VACUUM TUBE HANDLING AND USAGE PRECAUTIONS

There are still enough tubes in use, even in new equipment, to be worth listing rules for safe use.

- Tubes operate on dangerously high voltages. Always assume that every terminal of an operating tube is dangerous until you've measured the actual voltage.

- The capacitors in equipment store voltage, even after power is turned off, and they can recharge by themselves. In tubed equipment, these voltages are dangerous. Discharge all capacitors immediately before working on equipment.
- Tubes get hot enough to burn you badly. Some get hot enough to light paper on fire. Also, anything near a tube is likely to get very hot.
- If the glass tube breaks, that leaves the plate exposed, which is the highest-voltage part of a tube.
- Tubes get weak slowly, over a period of years, so slowly you don't notice. When servicing tubed equipement, take each tube out and put it in a tube tester to verify that it's okay. If you can't find a tube tester, measure voltages in the circuit to make sure they are what they're supposed to be.
- The electrical connections between tube pins and socket tend to get crusty. When replacing tubes, it's a good idea to spray the contacts with an electronic contact cleaner. (WD-40 is not a contact cleaner.) Put the new tube into the socket and wiggle it around a bit to clean the contacts.
- A tube's plate (anode) should not glow red hot. If it does, the tube is drawing excessive current. This may be due to a tube malfunction or malfunction of some other part. Check circuit voltages.

And that's all I'm going to say about tubes, aside from brief mentions in EM5 and EM6, because they've been almost completely replaced by transistors. If your career takes you to a place where you're actually dealing with vacuum tubes, you'll be getting specialized training for that.

The main thing you need to take with you, going forward, is the concept of the triode as an **electronically controlled current valve**. This basic concept also applies to the transistor, because it is also a 3-element device (though it's not called a triode) that's an electronically controlled current valve.

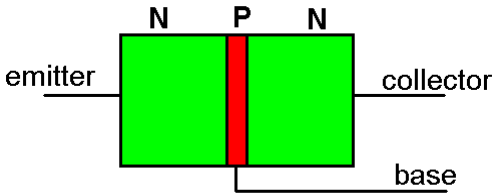
TRANSISTORS

There are two types of transistors:

- the **bipolar junction transistor** or **BJT**, which uses a **small current** to control a **large current**
- the **field-effect transistor** or **FET**, which uses a **voltage** to control a **current**

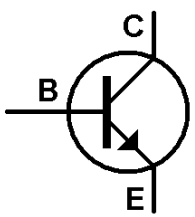
Transistors can operate at much lower voltages than tubes. In fact, as we will see, the voltage across a transistor often goes almost to zero.

Bipolar Transistors



The first kind of transistor, and the only kind from its invention in the late 1940s through the mid-1960s, was the **bipolar junction transistor** or **BJT**. A detailed discussion of a BJT's internal operation would be pretty complex, and there's no reason for you to care, because you're never going to work on the insides of a transistor. The important thing for a tech to know is how a transistor behaves in a circuit.

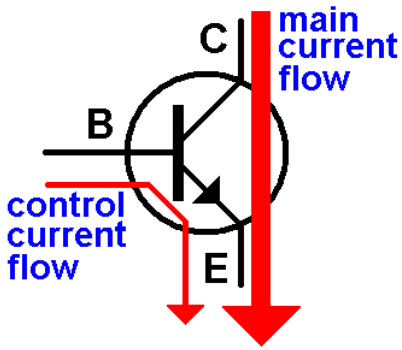
The most common type of BJT can be thought of as a semiconductor sandwich consisting of two N-type regions with a thin P-type region in the middle; thus, there are two back-to-back PN junctions. A lead is connected to each of the three sections. This is called an **NPN** transistor. As shown in the schematic symbol at left, the three terminals are called **emitter**, **base** and **collector**, abbreviated **E**, **B** and **C**.



- The emitter serves the same function as the triode's cathode; it is the source of electrons that flow from the negative DC power supply rail into the transistor.
- The base has the same function as the grid; it's the part that controls current flow.
- The collector is equivalent to the plate; it's normally at the most positive voltage and it collects the electrons that came from the emitter.

With tubes, you pretty much have to think in terms of electron flow; tubes don't make sense if you're thinking in conventional current. However, with transistors, even though the terminal names suggest electron flow, most engineers think in conventional current terms, and that's how I'm going to describe transistor operation.

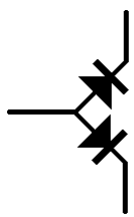
The main current path is the collector-emitter circuit. Note that in the schematic symbol, as with the diode, the emitter arrow points in the direction of conventional current flow. In a BJT, the base is the control input.



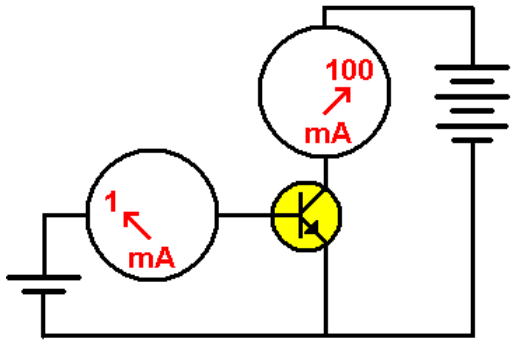
There are two key differences between a triode and a BJT.

- A triode is a *voltage-controlled* valve. A BJT is a *current-controlled* valve.
- The triode is normally "on" – it normally conducts current if a positive plate voltage is present. You have to apply negative voltage at the grid to turn it off. A BJT is normally "off" – you have to make current flow through the base to turn it on. And how do you make base current flow? You raise the base voltage about 0.7V above the emitter voltage, forward-biasing the base-emitter junction.

Since a transistor is two back-to-back junctions, in some ways it resembles back-to-back diodes. However, if you connect two diodes like this and then forward-bias the lower diode, the upper diode remains reverse-biased, and no current flows through it. Two regular diodes do not make a transistor.



But if we make the region that is common to both junctions, the base, very thin, that changes things. If you forward-bias the base-emitter junction, the collector circuit opens up and current flows. Not just some current, but a lot of current, much more current than is flowing through the base. In a typical small transistor, collector current will be 100 times base current or more.



It's not very useful to measure a BJT's gain in terms of transconductance. Instead, we measure **current gain**, the ratio of collector current to base current. Current gain is denoted by the Greek letter Beta, β. The BJT above has a beta of 100.

We can put this into a simple formula:

$$I_C = \beta \times I_B$$

where I_C is collector current and I_B is base current.

One of the most significant things about that formula is what is not in it: you don't see any mention of collector-emitter voltage. According to the equation, collector-emitter current is not affected by collector-emitter voltage. For example, suppose we have a transistor with a beta of 100, and we connect 3V across collector and emitter. If we put 0.1mA into the base, how much current will flow in the collector-emitter circuit? 10mA, right. Now let's increase that 3V to 15V. How much current will flow? Still 10mA. So current is independent of voltage.

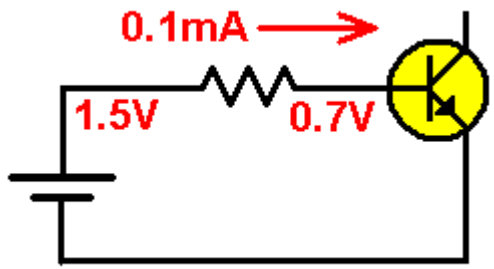
You probably know me well enough by now to suspect that I'm being a bit sloppy. Beta does in fact vary a bit with voltage, and also with current and temperature. But it varies much more from one transistor to the next, even when they're both stamped with the same part number. That's just a fact of life with transistors: it's hard for the manufacturer to control beta. The only way to get a batch of transistors with the same gain is to sort them.

This being the case, it's impossible to calculate a real-world transistor's collector-emitter current precisely, so there's no point in trying. You might as well pretend that the simple equation $I_c = B \times I_b$ is accurate, as most engineers do, and as I'm going to do.

Let's look at how we set the base current. We already know the base-emitter junction is a diode, so in normal operation the base will be 0.7V above the emitter. That's one of the fundamentals of transistor circuits, which you might as well commit to memory: when the transistor is “on,” which is to say that (conventional) current is flowing from collector to emitter, the base-to-emitter voltage will be about 0.7V, or a bit less at low current levels. Of course, we're talking about silicon. Transistors made of germanium have a 0.2 to 0.3V forward dropping voltage and some other exotic materials have different base-emitter dropping voltages. You won't run into them often. It's convenient that the base-emitter voltage is predetermined because it's one less variable we have to consider.

Now we have to control the base current. As you remember, a very small increase in the voltage across a diode causes a huge increase in current. Therefore, it's difficult to control base current by controlling base voltage. Moreover, if the transistor heats up, the base-emitter voltage will drop a little, causing more base current to flow, causing more collector current to flow, causing further heating; a problem that continually makes itself worse. This is called **thermal runaway**. Preventing thermal runaway is a large part of the design task an engineer faces when designing high-power circuits.

So, since controlling base current by controlling base-emitter voltage doesn't work very well, we have to control the current directly.



The easiest way to do this is to insert a resistor in the circuit, and calculate voltage drops and resistance to get the base current we want. We already know the base-emitter voltage, 0.7 volts, so this is a simple Ohm's Law calculation.

Let's say we want 10mA of collector current, and our transistor has a beta of 100, so we want 0.1mA to flow in the base. Suppose we're using a 1.5V battery to supply the base current. We just need to insert a dropping resistor between the battery and the base.

$$\mathbf{1.5V - 0.7V = 0.8V \text{ across resistor}}$$
$$\mathbf{0.8V / 10^{-4}A = 0.8V \times 10^{4} = 8k\Omega}$$

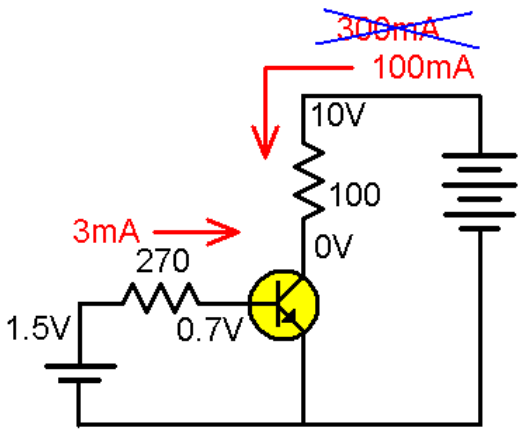
In the real world, we'd use the closest standard value, which is 8.2k. And that, ladies and gentlemen, is how we control collector current. As long as the voltage at the collector is positive with respect to the emitter, and as long as we know

the transistor's beta, we can (within limits) get any collector current flow we want by providing the proper base current.

BJTs as switches

As we've seen, collector current is equal to base current times beta. If there's a resistor in the collector circuit, there will be a voltage drop across it, which is determined by collector current and resistance, V = IxR. But what if the resistance is so high that there isn't enough current available to equal I_B x β?

Suppose, for example, that we have a 10V supply, a 100-Ohm collector resistor, a base current of 3mA and a β of 100? 3mA x 100 is 0.3A. So, the transistor would pass 0.3A if there were that much current available.



But there isn't. The maximum current that can flow in this circuit, if the transistor turns all the way "on" and puts the whole 10V across the resistor, is 10V / 100Ω = 0.1A.

So that's what happens. The transistor pulls the lower end of the resistor all the way down to zero volts (in the real world, *almost* zero volts), and all 10V is dropped across the resistor. In other words, the transistor is the equivalent of a closed switch.

This condition, when a transistor turns fully "on" and the voltage across it falls to almost zero, is called **saturation** – the same word we use with magnetic materials, but with a different meaning.

To make a transistor work as a switch, we simply have to drive the base hard enough to turn the transistor on enough to pull the collector down to zero volts.

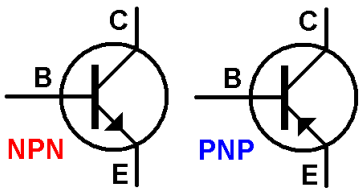
A transistor in saturation, i.e. fully on, takes longer to turn off than it took to turn on. Think of switch contacts with sticky goo on them; once they're pushed together, it takes some time to pull them apart. In circuits where transistors have to switch on and off quickly, that's a significant problem. We'll be talking more about that.

Transistors have replaced mechanical switches in a huge number of applications. They have several important advantages:

- They don't wear out. Assuming the circuit is designed so none of the transistor's limits are exceeded, it should work forever.
- They can be directly controlled by electronic circuitry, so we can do things like use a computer to control the motors in a machine.
- They can switch much faster. Some transistors can switch at GHz speeds.

BJT POLARITY

Up to this point, we've been looking only at NPN transistors with, in order, N, P and N doping. It is also possible to make transistors the other way, with P, N and P doping. These are



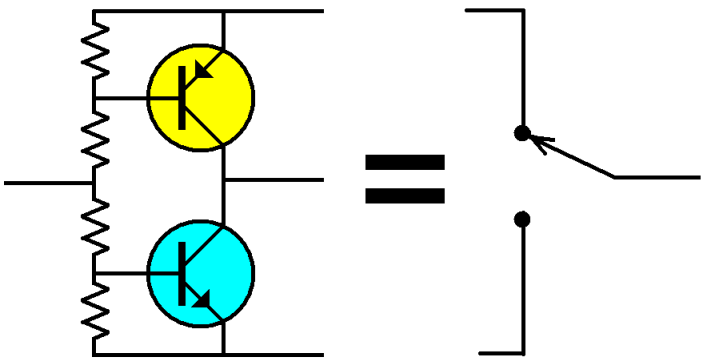
called **PNP** transistors. (Try to hide your astonishment.) Note that the arrow points the other way, denoting the direction of conventional current flow in the emitter-base PN junction.

A PNP transistor is the mirror image or **complement** of an NPN. Everything I've said about NPN transistors also applies to PNP transistors, but with all polarities reversed. The emitter is the most positive terminal, the base is less positive than the emitter by -0.7V, and the collector is the most negative. We could examine how electrons flow in a PNP transistor, but it's easier to just think of positively charged holes and conventional current flow.

By the way, “compliment,” with an i, means saying something nice. “Complement,” with an e, means something that's the corresponding opposite of something else, like a PNP transistor to an NPN. Why am I putting a spelling lesson into an electronics tute? Mom was an English teacher. Sue me.

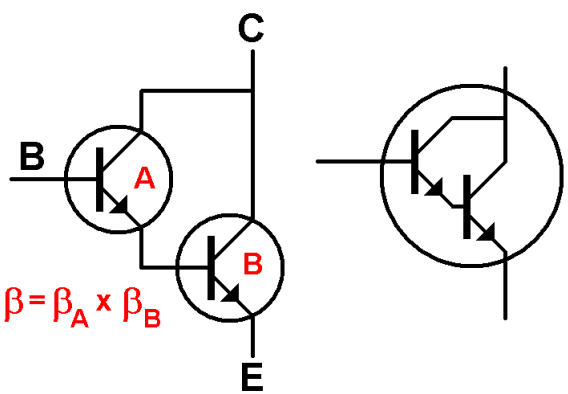
Complementary transistors permit us to do things that can't be done with tubes. For example, if we flip over the PNP transistor so its emitter goes toward the positive rail (the usual case) we can make a circuit with a PNP to pull a terminal up to the positive supply rail, and an NPN to pull it down to the negative rail.

In other words, an SPDT switch. (I'm assuming DC voltages here. The circuit wouldn't work well in an AC environment.) I'll show you more examples in the next chapter.



DARLINGTON TRANSISTORS

If you need more current gain than you can get from one transistor, you can always add a second one. Probably the simplest example is the **Darlington** circuit, which can easily have a beta of 10,000 or more. It can be built with two separate transistors, but it's so widely used that transistor manufacturers put both transistors in a single 3-terminal case, so engineers can treat it almost like a regular transistor.



The main differences from regular transistors, aside from high gain, are that the base-emitter drop is about 1.4V (because there are two junctions in series) and that Darlingtons tend to be slower than single transistors.

HETEROJUNCTION BJTs

In heterojunction transistors, the base and emitter are made of different materials. Their great advantage is high speed. It's difficult to get normal BJTs to operate beyond about 4GHz, but heterojunction transistors can operate at hundreds of GHz. They're exotic today, but you'll run into them more often in the future as we learn how to make use of those extremely high frequencies.

SILICON CARBIDE TRANSISTORS

Silicon carbide transistors work pretty much like regular transistors, except that they can handle higher temperatures. They're beginning to become common in high-power control circuits.

BJT OPERATION SUMMARY

The basic operation of a BJT is really simple:

- A BJT is a current valve.
- Current flow is current-controlled.
- Its parts are the emitter, which is the source of electrons, the collector, which is their destination, and the base, which is the control element.
- Conventional current flow is from collector to emitter.
- A BJT is normally “off.” To turn it on, apply a positive current to the base. (A negative current in the case of a PNP transistor.)
- A BJT can operate at low voltages. However, kilovolt BJT's are available.

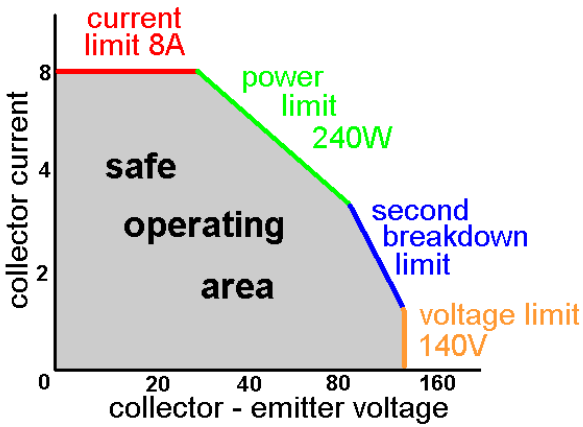
BJT HANDLING AND USAGE PRECAUTIONS

A transistor will be destroyed if any of the following limits are exceeded:

- Maximum power dissipation: collector-emitter voltage times collector-emitter current.
- maximum collector current
- maximum collector-emitter voltage
- maximum base current
- second breakdown

Second breakdown is a phenomenon where current is not distributed evenly, but concentrates itself in a small area. It's most prevalent at high voltages and high frequencies. Second breakdown can destroy a transistor in microseconds, which is why you'll often find complicated protection circuitry in high-powered equipment.

All these factors except base current can be combined in a **SOA** or **safe operating area** graph which identifies maximum voltage and current under all conditions. Engineers use these graphs to make sure the transistor is never stressed beyond its limits.



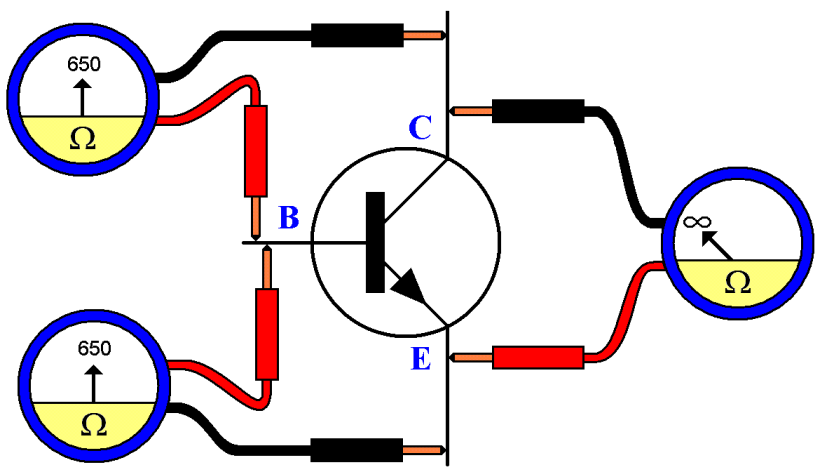
The SOA graph above is for a typical 140V, 8A, 240W transistor. Note that each of the limit quantities, plus second breakdown, governs the SOA in different parts of the graph. Note that both X and Y axes are logarithmic. This permits the power-related segments of the SOA boundary to be drawn as straight lines.

If the SOA limits are exceeded in any direction for even microseconds, the transistor may be destroyed. In high-power circuits where the limits may be approached, you will see elaborate protection circuitry to keep this from happening.

One more precaution: molten solder can melt the silicon in a transistor just like a diode. Use care when soldering.

TESTING A BJT

Solid state devices are usually either 100% good or 100% dead. Since a BJT has two back-to-back diode junctions, you can test it with an ohmmeter.



If it's an NPN transistor, the base-emitter and base-collector junctions will conduct if you put the red lead on the base, but will not conduct if you put the black lead on the base. For a PNP transistor, the reverse is true. The collector-emitter junction should not conduct no matter which way you connect the leads. If any of these conditions are not met, the transistor is bad.

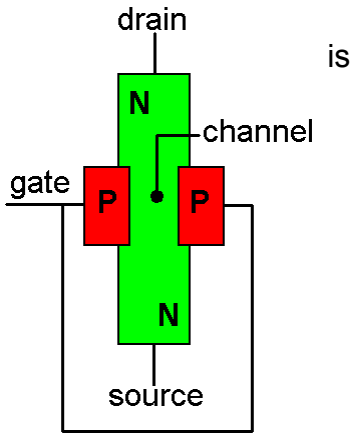
Many VOMs have sockets that let you plug in a small-signal transistor; they'll not only tell you if the transistor is good, but measure its gain. You do have to know something about the transistor before you start – whether it's NPN or PNP, whether the lead pattern is EBC or ECB, and of course, whether it's a BJT at all, or an FET.

FIELD EFFECT TRANSISTORS

Another kind of transistor began to emerge in the 1960s, the **FET** or **field-effect transistor**. The earliest kind was the **junction FET** or **JFET**. Like BJT's, there are two complementary types, **N-channel** and **P-channel** JFETs. The big difference from a BJT is how current is controlled. A BJT is a current-controlled device, while an FET is a voltage-controlled device.

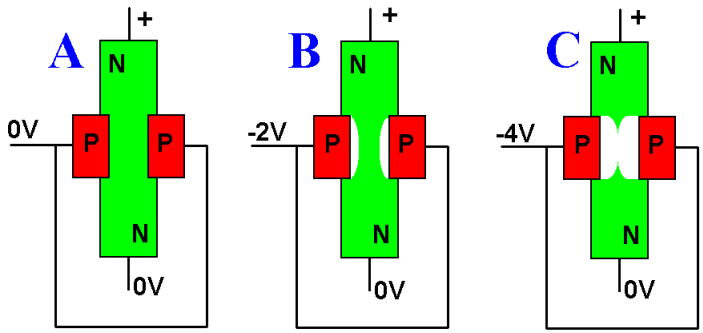
JUNCTION FETs

An N-JFET consists of a continuous piece of silicon which is N-doped to make it a conductor. Electrons enter the silicon at the **source** and exit the **drain**. Between source and drain, electrons must flow through a narrowed area called the **channel**, which is bounded by P-doped silicon called the **gate**. This is literally a gate; it opens or closes the channel to control electron flow. (Finally, a device where all three terminals have names that describe their functions!)



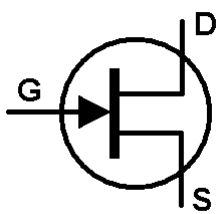
If we were to put a positive voltage on the gate, we would create a forward-biased junction and electrons would flow from the N-section into the P-section. However, if we put a negative voltage on it, creating a reverse-biased junction, an interesting thing happens. Since the channel-gate junction is now reverse biased, no current flows in the gate.

However, the negative charge in the P-region repels the electrons flowing through the channel. This effectively reduces the width of the current-carrying part of the channel, like stepping on a garden hose, and the FET begins to turn "off." At "A" in the illustration below, gate voltage (relative to the source) is 0V and current can flow through the channel. At "B" the gate is -2V and current flow is starting to be restricted.

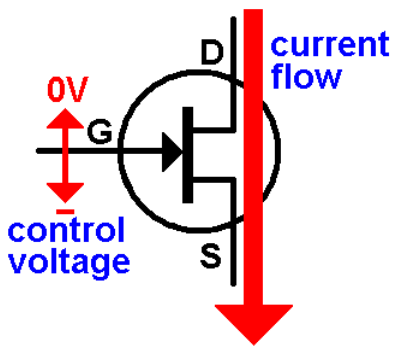


If we make the voltage in the P-region negative enough, the current-carrying part of the channel will shrink to zero and current flow will be stopped completely as at "C" above. This is called **pinch-off**.

Here's the symbol with source, gate and drain labeled. The gate-to-channel junction is a diode, and the gate is P-doped, but in normal operation this junction is reverse-biased so no current flows in the gate. P-channel FETs are also available; in the symbol, the arrow points the other way.



The source and drain look the same, so how do we tell which is which? Well, in theory there is no difference. Current ought to be able to flow through the FET either way. A few FET's have been made that are perfectly symmetrical, intended for bidirectional current flow. Most of the time, though, there are differences. You just have to follow the manufacturer's specification. If they say a particular lead is the source, then it is.



Unlike a BJT, current is controlled by voltage, specifically voltage between gate and emitter. A JFET is normally "on," and the gate voltage must be made negative to reduce or stop current flow. In this way it's like a triode. And like a triode, its most basic characteristic, sensitivity to gate voltage changes, is specified as a transconductance. Just for reference, 1 mho or 1 Siemens of transconductance means that a 1V change in gate voltage results in a 1A change in drain current. Real-world JFETs have transconductance of 0.001S (1mS, which in this case is milli-Siemens, not millisecond) to 0.03S (30 mS).

The problem with using FETs for amplifiers is that, while tube behavior is pretty uniform from one tube to the next, FETs vary all over the place. The gate voltage necessary to achieve pinch-off may be -1V for one FET and -4V for the next. There are ways of handling that, which you will learn in chapter 3.

Like a triode's grid, the FET's gate is an open circuit at low frequencies. At higher frequencies, drive current is needed to charge and discharge the various capacitances involved. Nevertheless, depending on many factors, drive current is often much lower than with BJT's. JFET's were used as high frequency amplifiers at a time when high-frequency BJTs were hard to make. They have naturally low noise and make good low-level amplifiers, especially with high-impedance signal sources.

JFET OPERATION SUMMARY

- A JFET is a current valve.
- Current flow is voltage-controlled.
- Its parts are the source, which is the source of electrons, the drain, which is their destination, and the gate, which is the control element.
- Electron flow is from source to drain. Conventional current flow is from drain to source.
- A JFET is normally "on." To turn an N-channel JFET off, apply a negative voltage to the gate.

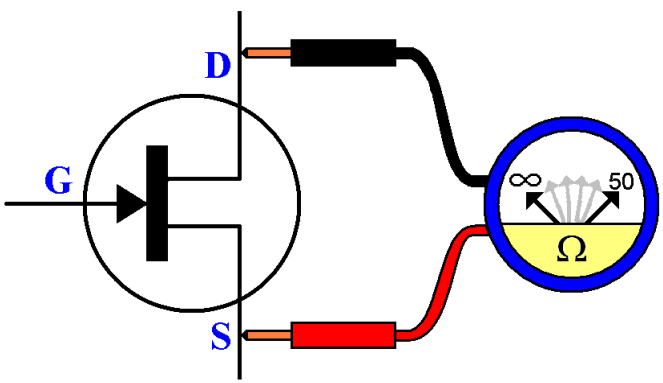
JFET HANDLING AND USAGE PRECAUTIONS

- JFETs have the same maximum voltage, current and power restrictions as BJTs.
- JFETs are not overly static sensitive, but you should avoid a hot discharge into the circuit.
- Be careful when soldering.

TESTING A JFET

This is going to drive some people crazy. It involves your fingers. The first thing is to connect the ohmmeter across the source and drain. In theory, it shouldn't matter which lead goes where, because the FET is essentially a resistor when the voltage across its leads is low.

Now put a finger across the source and the gate. This should discharge any voltage difference, putting the gate at the same voltage as the source. The ohmmeter should tell you that the resistance across source and drain is very low.



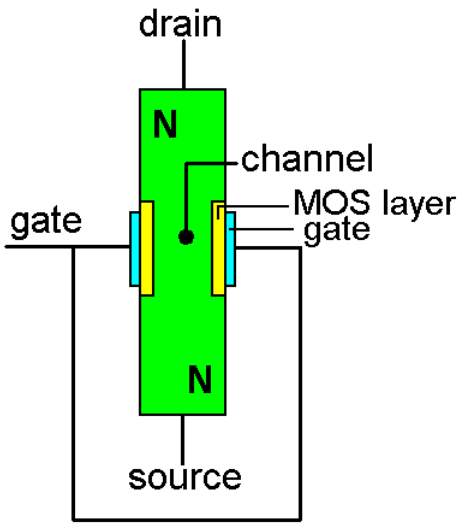
Next, let go of the source and touch the gate several times. The ohmmeter should fluctuate, giving various readings between a low resistance and an open circuit. What you're doing is putting random amounts of charge on the gate, and the fluctuating meter readings tell you the FET is working normally.

As I write this, I'm sitting at home in Minnesota and I can hear the screams of outrage all the way from New Zealand. Look, folks, this is a junction FET. Any overvoltage at the gate is either going to forward bias the gate-channel diode, or reverse bias enough to put it in zener mode. Either way, it will be discharged. This is exactly what would happen with a BJT, and nobody thinks of BJTs as static-sensitive.

MOSFETS

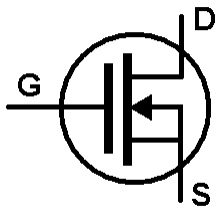
The first thing you need to know about MOSFETs is that you do *not* use the finger-on-the-gate method to test them. MOSFETs are very static sensitive. Just pushing the leads into a piece of styrofoam can create enough static to ruin a MOSFET. When you're working with MOSFETs, you and your soldering iron and ideally your metal benchtop should be Earth grounded through 1MΩ resistors. The 1MΩ resistors are there so if you happen to get yourself connected to the AC power line, it won't result in enough current to harm you.

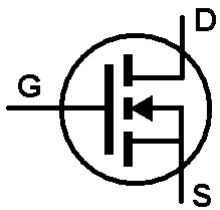
"MOSFET" stands for **metal oxide semiconductor FET.** In a MOSFET, instead of a doped gate, the gate is a simple piece of metal plating. The gate is insulated from the channel by a thin layer of nonconductive metal oxide or similar material. Since we now have a gate that cannot conduct under any conditions (unless the MOS breakdown voltage is exceeded), it is no longer necessary to keep the FET reverse-biased. This permits us to make "normally on" and "normally off" FETs.



A "normally on" NMOSFET (N-channel MOSFET) works like a regular N-channel JFET: when the gate voltage is zero, relative to the source, the channel conducts and the FET is "on." To turn it "off," we put a negative voltage on the gate. A "normally off" NMOS-FET works the opposite way: when the gate voltage is zero, the channel does not conduct, so the FET is "off." To turn it on, we put a *positive* voltage on the gate. In this sense it's a little like a BJT, except that a FET uses a voltage rather than a current to turn the unit on, and there is no gate current.

A normally on MOSFET is called a **depletion mode** FET because controlling the current – reducing it or turning it off – involves depleting the channel's conduction zone of electrons. In the schematic symbol, the channel is shown as a solid bar to indicate that it normally conducts.

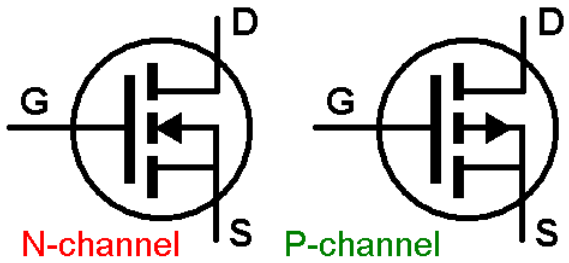




A normally off MOSFET is called an **enhancement mode** FET because controlling the current – turning it on – involves enhancing the conductivity of the channel, creating a positive field in the channel to encourage electron flow. In the schematic symbol, the channel is broken to show that it normally does not conduct. (There are no enhancement-mode JFETs, because applying a positive voltage to the gate would forward-bias the junction.)

The symbols above are for N-channel FETs. It may seem that the arrow is going the wrong direction, but in a MOSFET, there is no gate-to-source junction. The arrow actually shows the junction between the channel and the substrate, called a **parasitic diode**. Sometimes that diode is use to snub flybacks; otherwise it doesn't have much to do with circuit operation.

P-CHANNEL FETS



Like BJTs, both JFETs and MOSFETs are available in N-channel and P-channel versions. From the standpoint of usage, low-power N- and P-FETS are electrically identical, mirror images of each other. At high power levels, there are some differences. It's easier to make a brawny NMOSFET than a brawny PMOSFET, so the PMOS versions are more expensive. Engineers tend to design with NMOS when they can.

In high-power switching applications, MOSFETs are extremely attractive because the best ones have very low on-resistance, on the order of thousandths of an ohm. For an old-school guy like me, the ability to use a flea-powered control signal to switch hundreds of amps, as some MOSFETs can do, is mind-boggling.

By the way, even though gate insulating layers are not necessarily metal oxide these days, “MOSFET” is still the most commonly used name. A more general name is **IGFET** for **insulated gate FET**. Most people don't use that name because it's too easy to confuse with IGBT, another type of transistor that we'll study next.

MOSFET OPERATION SUMMARY

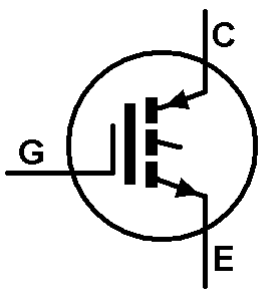
- A MOSFET is a current valve.
- Current flow is voltage-controlled.
- Its parts are the source, which is the source of electrons, the drain, which is their destination, and the gate, which is the control element.
- Conventional current flow is from drain to source.
- A depletion-mode MOSFET is normally “on.” To turn it off, apply a negative voltage to the gate.
- An enhancement-mode MOSFET is normally “off.” To turn it on, apply a positive voltage to the gate.

MOSFET HANDLING AND USAGE PRECAUTIONS

- Low-power MOSFETs are very easily destroyed by a gate-to-channel voltage of either polarity that's more than the insulating layer can withstand. They are less vulnerable when they're actually in a circuit, because other components tend to drain off charges. Never touch a finger or tool to the gate until you've touched the source or the chassis of the unit. When the MOSFET is not wired into a circuit, keep the leads shorted together.

- High power MOSFETs are less sensitive, but you should still observe the cautions listed above, if maybe a little less carefully.
- When you get FETs from an electronic supplier, they will either come in a gray plastic bag or their leads will be stuck into black foam, or both. Don't throw these things away, they're useful. They are special bags and foam that are slightly conductive. While FETs are stored in them, they're protected from static discharge, and you can handle them like you'd handle any other transistor.
- Be careful when soldering.

IGBTs

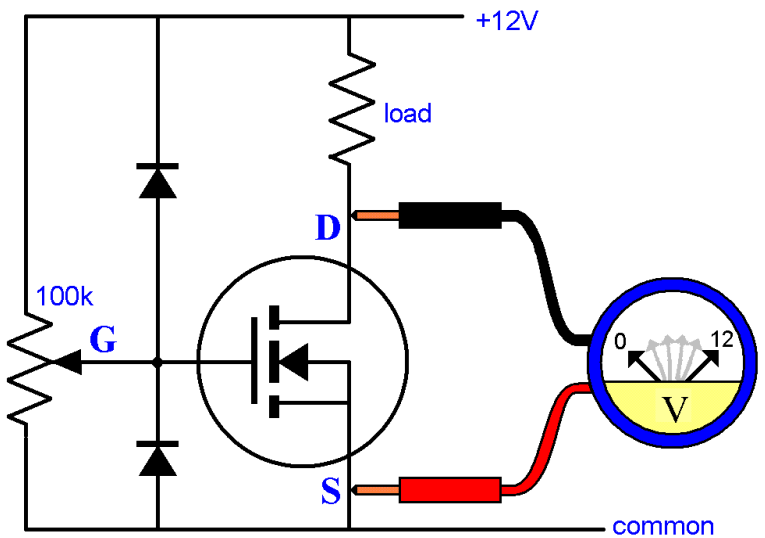


An **insulated-gate bipolar transistor**, abbreviated **IGBT**, is a hybrid transistor with a BJT's collector-circuit behavior but an open-circuit input, like a FET. The terminal names also borrow from both: emitter, gate, collector.

The IGBT is the gorilla of control devices. IGBT's are made in modules up to the size of small bricks that can switch thousands of volts at hundreds of amps. Connections are often via screw terminals. They're used in electric cars and trolleys, power plants and all kinds of big machinery.

TESTING MOSFETS AND IGBTS

Due to the sensitive gate, the only good way to test a MOSFET or IGBT is to make a test jig that puts a voltage across it, with a pot to adjust gate voltage and diodes to drain off static. When a MOSFET or IGBT is connected into the jig, it's protected against static discharge.



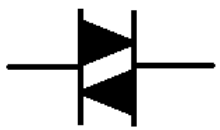
For high-power devices, a car brake light makes a good load, and then you don't need the meter. You won't have to provide a bipolar supply because almost all MOSFETs and IGBTs are enhancement mode.

5. AVALANCHE DEVICES

A transistor is a valve; it can be on, off, or partially on. It can be made to work like a switch, but that's not a built-in characteristic.

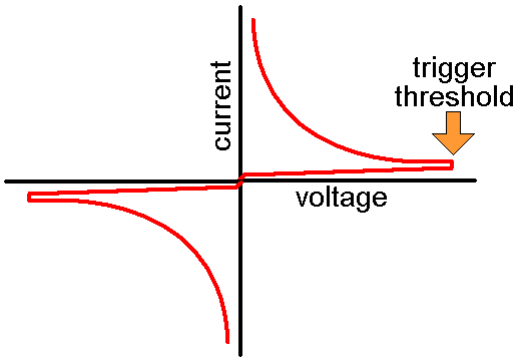
An avalanche device is a switch; it's either off or on, open or nearly shorted, and the transition is sharp and virtually instantaneous. It cannot be made to turn on slowly, or turn partway on. It's called an avalanche device because, once it starts conducting, current snowballs and goes right to the max.

TRIGGER DIODES



There are unidirectional trigger diodes, which only conduct one way, and bidirectional trigger diodes, which can conduct either way. I'm going to ignore the obvious question (whether a bidirectional device can truly be a diode) and call them all "diodes" as most people do.

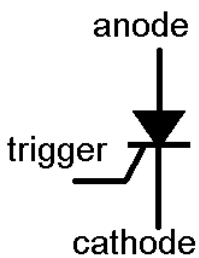
If voltage across a trigger diode is gradually increased, it will not conduct at all (except for leakage) until the trigger voltage is reached; then it turns completely "on." As shown in the graph above right, voltage across the diode immediately falls to near zero once the trigger voltage is reached. That's because it becomes almost a short circuit, like a closed switch. Most trigger diodes, like the one shown, will trigger in either direction. Another name for a bidirectional trigger diode is a **diac**.



Once triggered, the diac will conduct as long as current flows; there is no way to turn the diac "off" except by removing the source of current. This makes trigger diodes nearly useless in DC circuits. However, in AC circuits the current reverses 120 times a second, so there are two moments in each cycle when current naturally falls to zero, called the **zero crossings**. This takes care of the turn-off problem.

The most common trigger voltage is around 30 volts, though diodes that trigger at other voltages are available.

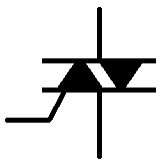
SCR's



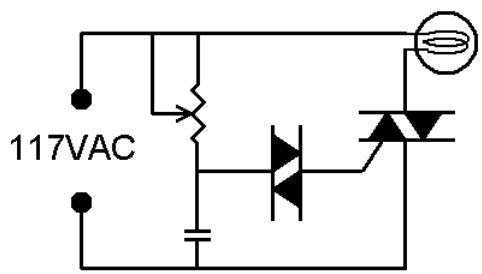
SCR stands for **silicon controlled rectifier**, also called a **thyristor**. Like other rectifiers, an SCR will only conduct one way. An SCR has an anode, a cathode, and a trigger input.

Current flow is zero until a positive voltage is applied to the trigger, turning the SCR on. It's like a regular diode that conducts in the forward direction, except that it has to wait for “permission” in the form of a current pulse at the trigger terminal. Like a trigger diode, once it's on, it stays on until current disappears. Thus the trigger current only has to last an instant.

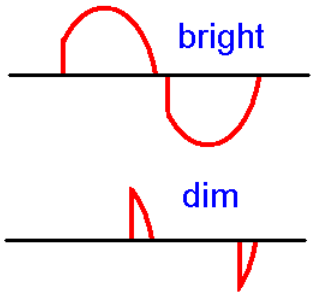
TRIACS



A triac is like a pair of SCRs connected together so as to conduct current either way. It is turned on by applying a voltage to the trigger input of the same polarity as the voltage across the triac – in AC applications, a positive trigger voltage during the positive alternation and a negative trigger voltage during the negative alternation.

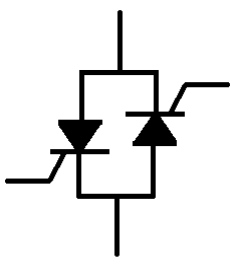


The easiest way to do this is to charge up a capacitor through a resistor, then apply the charge to the trigger input through a trigger diode. If the resistor is changed, that will change how long it takes the capacitor to charge to the trigger voltage. This is essentially the circuit used by wall-type incandescent light dimmers. The triac doesn't turn on until partway into each half cycle, then it turns off when the AC waveform crosses the zero-voltage point. If it turns on early in each half-cycle, the lamp will glow brightly. If it waits until each half-cycle is almost over, the lamp will be dim.



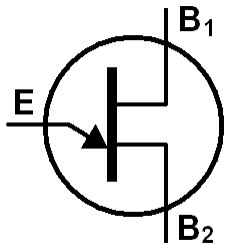
Because SCR's and triacs turn on abruptly, the current waveform has virtually instantaneous rise times which generate a lot of RF energy. Real-world dimmers have to include filter components to slow down the wavefronts and suppress noise.

Special triacs have been developed that can be turned off by applying an opposite-polarity voltage to the trigger terminal. These are relatively rare as FET's are taking over that function.



Often, in a very high-power circuit, a pair of SCRs is used instead of a triac. This allows the power dissipation to be shared between two devices, increasing maximum power handling capability.

UNIJUNCTION TRANSISTORS



The **UJT,** or unijunction transistor, is completely obsolete; they haven't been made in decades. However, you will still run across them in older equipment. A manufacturer won't throw away a $200,000 machine when it gets to be a few years old, they way people do with their cellphones. You could very well find yourself repairing equipment made before you were born.

UJTs were popular in the 1960s and 70s because they provided a simple way to generate a continuous string of pulses with a single device. That's one kind of **oscillator**, and we have a whole chapter on oscillators coming up, so I won't discuss the UJT any more here. There's a similar device called a PUT or programmable unijunction transistor, which does the same thing but allows you to choose the tripping point with a couple of resistors. PUTs are still being made and stocked by electronics distributors.

UJTs are becoming impossible to find. If you need to replace one, you're probably better off modifying the circuit to accept a PUT.

AVALANCHE DEVICE OPERATION SUMMARY

- An avalanche device turns on all at once. It cannot be turned partway on.
- Once turned on, the only way to turn most avalanche devices off is to remove the source of current.

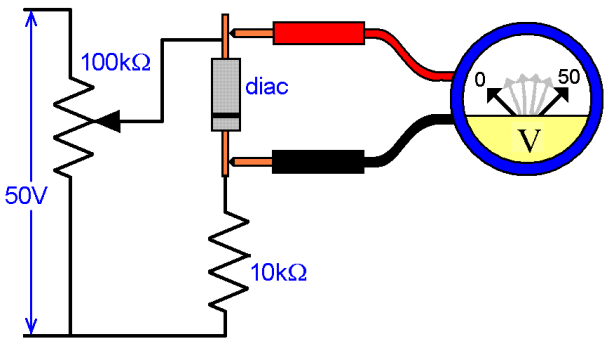
AVALANCHE DEVICE USAGE AND PRECAUTIONS

- Manufacturers' maximum peak-current specs for trigger diodes and UJTs should be observed. The timing circuit usually involves a capacitor between the device and common. When it triggers, the capacitor is almost instantly discharged through the device. If too large a capacitor is used, the device can be destroyed.
- SCRs and triacs are almost always used in AC line-connected circuits. That means they're dangerous. It also means that the trigger circuits connected to them are dangerous. **You can't assume the low-power parts of the circuit are safe just because they're low-power!** Proceed as if the whole circuit is electrically “hot” unless you are certain that there's isolation (e.g. transformer or opto-isolator) involved.
- Observe the usual cautions when soldering. Avalanche devices are usually not static sensitive.

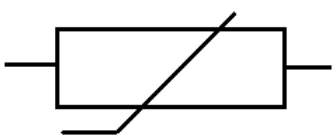
TESTING AVALANCHE DEVICES

You can always swap a questionable device for a new one, but that means you need to have a new one in hand. What you want, of course, is to test the existing one so you know whether you need to order a new one. You can't test an avalanche device with a VOM, you have invent a test circuit that uses the part in the way it's normally used.

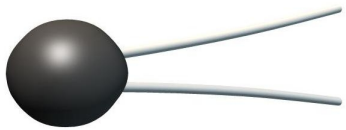
Here, for example, is a test circuit for a diac. As the voltage at the pot's tap rises, the meter reading should rise, telling you the diac isn't conducting. When its trigger voltage is reached, the meter should go to almost zero.



6. THERMISTORS

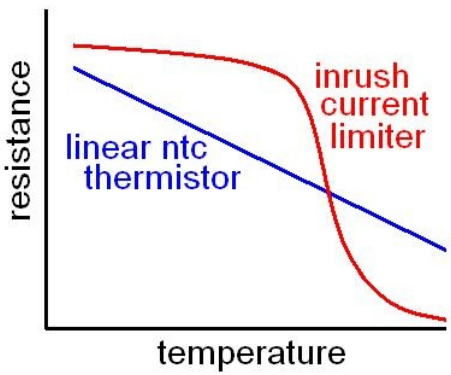


We could really get bogged down over the definition of "active device." I've included the thermistors in this chapter because you should know about them, because they are solid state devices more or less based on semiconductor materials, and because I can't think of a better place to put them.



A thermistor is a device whose resistance varies with temperature. In a **positive temperature coefficient** or **PTC** thermistor, resistance rises with rising temperature. In a **negative temperature coefficient** or **NTC** thermistor, resistance falls with rising temperature.

In most thermistors, resistance varies somewhat linearly with temperature, especially over a small range. However, some are made to change little until the temperature reaches a certain point, then change very rapidly. One such type is the **inrush current limiter**. Many AC power line operated devices, e.g. anything with a big filter capacitor, draw a lot of current for an instant when first turned on. This is called **inrush**.



If an NTC thermistor is placed in series with the power line, its initially high resistance will reduce inrush current. Once current is flowing through it, it heats up and its resistance falls to the point where it essentially disappears from the circuit. Just any old thermistor won't work; it has to be one specifically designed for this use.

The opposite type, a PTC thermistor whose resistance rises rapidly when a certain temperature is reached, is used as a **self-resetting fuse**. If current exceeds an acceptable value, the thermistor heats up and its resistance greatly increases, reducing current. At that point, since there's a large voltage drop across the thermistor, it stays hot. When power is shut off, the thermistor cools down and its resistance returns to a low value, like a breaker that automatically resets when power is removed.

THERMISTOR SUMMARY

- A thermistor is a resistor whose value changes with temperature.
- A PTC thermistor's value rises when temperature rises.
- A NTC thermistor's value falls when temperature rises.
- A thermistor's resistance may vary gradually with temperature, or it may stay about the same until a threshold temperature is reached, then change a lot over a few degrees.
- A threshold-type NTC thermistor is typically used to limit inrush current when a piece of equipment is first turned on.
- A threshold-type PTC thermistor is typically used as a fuse that resets itself after a load fault is removed.

THERMISTOR USAGE PRECAUTIONS

Thermistors are **bulk devices**, meaning there are no fragile doped regions, they're just a lump of material, so they're fairly tough. However, like any measurement device, you have to be careful not to accidentally screw up the measurement.

- If a thermistor is used in a circuit that runs enough current through it to heat it, even a little, its purpose is defeated. It's much better to run a tiny bit of current (microamps) through it and then use a DC amplifier (chapter 4) to boost the output.
- If a thermistor is mounted on or near a heat-sinking object, that object may draw away the heat you're trying to measure. When you want to measure the temperature of an object, mount it so the thermistor body contacts that object as close to the heat source as possible. Use thin wires so they won't wick heat away from the thermistor. If you want to measure air temperature, mount the thermistor well away from

- surfaces and be sure the leads and mounting do not conduct heat to or away from the thermistor.
- Thermistors can measure the temperature of liquids, but then they have to be insulated. Industrial applications tend to use thermistors that are mounted in sturdy, sealed metal cases, available from a number of makers.

7. PACKAGES

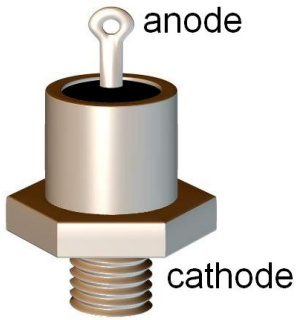
Standard semiconductor packages are identified by the letters “TO” followed by a number.

AXIAL PLASTIC DIODE

Diodes have long been made with axial-lead packages, like axial capacitors. These are prepared for insertion in PC boards by simply bending the leads. It is good practice to use a longnose pliers and bend the lead at some distance from the diode, rather than bending it against the diode case itself. It's especially important to do this with glass diodes; otherwise, the bending force can crack the glass.



STUD MOUNT RECTIFIER

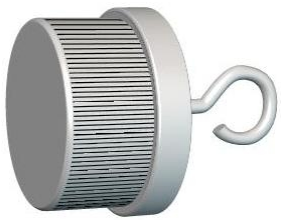


Rectifiers which carry high current must have a way of getting rid of heat. The forward voltage drop across a rectifier may be only one volt, but if it's carrying 25 amps, then that's 25 watts of heat to get rid of.

The traditional high-power diode package is a metal case with a threaded stud attached. This stud is used to attach the rectifier to a heat sink. The stud also serves as the cathode connection; the anode is a sturdy solder lug.

In some cases the polarity is reversed, to make it possible to chassis-ground the anode where necessary; be alert to this possibility. One manufacturer added an “R” to the part number stamped on the case to indicate reverse polarity. As far as I know, these aren't being made any more, but they were very popular in their time. Keep this in mind in case you run across them in older equipment.

Pressfit rectifier

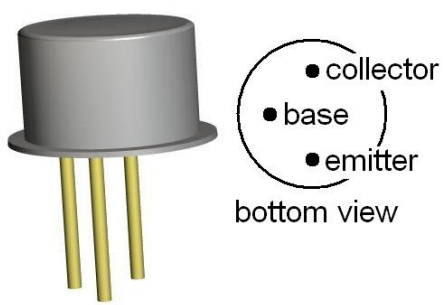


A rectifier designed to be pressed into a hole in a metal heatsink. Grooves on the body keep the rectifier in place and insure good contact. The case can be the anode or cathode. Widely used in automotive alternators.

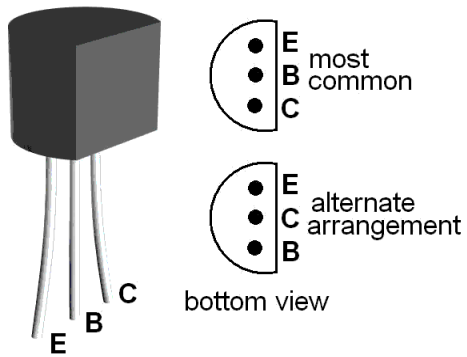
Low Power Transistors

Low-power transistors are also called **small signal** transistors, a somewhat archaic term meaning that the signal voltages and currents normally handled by the transistor are small compared to the DC voltages and currents in the circuit. For example, in a circuit where the DC power supply is 30V but the largest AC signal level in the circuit is 30mV, that would be a “small signal” and a transistor designed for that application would be a “small signal transistor.” Nowadays the expression means a small transistor, i.e. one whose maximum power rating is below a watt.

Older transistors are encased in a metal can about ¼ inch in diameter. This is called the **TO-5 case**. If you look at the bottom, the leads form a triangle and their positions correspond to the way an NPN transistor is most often drawn in a schematic diagram: base to the left, emitter down, collector up.



Most small transistors are now encased in plastic. The leads can either form a triangle or be arranged in a line. The most common configuration, from left to right looking at the flat side, is emitter, base, collector. But do not assume this is the case. Some transistors are arranged emitter, collector, base.



These days, the easiest way to find out which is which is to go on the web and look up the part number. However, this doesn't always work. Many transistors have no printed part number or a “house number” that doesn't correspond to any standard designation. A transistor that's not wired into a circuit can be scoped out by taking an ohmmeter and measuring across the leads. For an NPN transistor, put the black test lead on the left transistor lead; that's always the emitter.

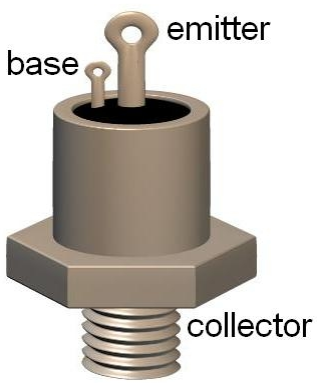
Then use the red test lead to find the one that conducts; that's the base. For a PNP transistor, reverse the test leads. If you don't know which kind it is, test both ways. Maybe you're lucky enough to have transistor test sockets and a transistor test function built into the meter, which makes things a lot simpler.

When the transistor is connected into the circuit, you may get strange test results due to the other components that are connected to it. And of course, if the transistor is defective, the measurements won't tell you anything. In this case, if the number isn't stamped on the part (or you can't read it because it's burned), you'll have to get the schematic and find out what that transistor is supposed to do.

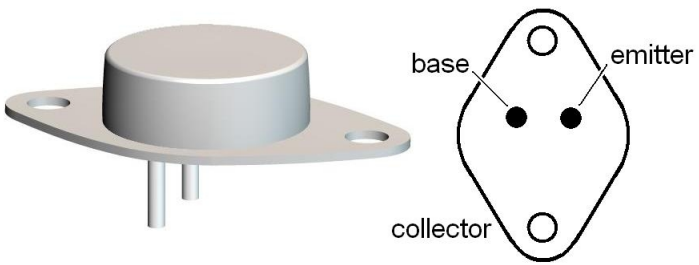
The little plastic transistor case is called a **TO-22**. That case size was first used for metal-cased transistors about $^{1}/_{8}$ inch in diameter, but now it refers to plastic transistors.

HIGH POWER TRANSISTORS

The original power transistor case was a stud rectifier case with a smaller solder lug added for the base. These are quite old, no newer than the 1960s; about the only place you find them now is in classic car radios. They're almost always germanium PNP transistors. Good luck finding replacements.



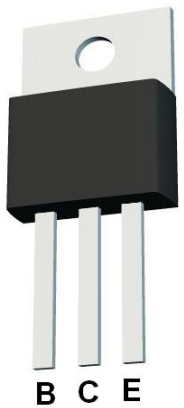
The next newer case was the **TO-3**, with pins for the emitter and base and a diamond-shaped plate about an inch wide for the collector. The transistor is mounted to a heat sink with machine screws and nuts that also serve as the collector connection.



If the collector has to be insulated from the heat sink, a thin mica washer is used to insulate the plate and plastic shoulder washers are used to insulate the mounting screws. This was the standard power transistor case for a long time, and you still occasionally find them, even in new equipment.

A little brother, the **TO-66** case, about half an inch wide, is now pretty much obsolete. If you ever need to replace one of these, the best bet is to substitute a newer, tab-type power transistor of the type described at right. The leads can be formed to fit where the TO-66 used to be.

At some point, people figured out that a metal heat sink surface could be molded right into a plastic case. The most common example is the **TO-220** case, which has three leads. From left to right: base, collector, emitter, or for a FET, gate, drain, source. The metal tab is connected to the collector/drain, and it's "legal" to make the electrical connection to the tab instead of the lead. Occasionally you'll even see the collector/drain lead cut off. The TO-220 is the most common power transistor case today. It has proved so practical that power diodes and even resistors are now made using the same case.



There are many variations on the TO-220 case:

- short tab without a screw hole (for soldering directly to a PC board)
- insulated case with a thin layer of plastic covering the heat tab, to eliminate the need for the mica washer and screw insulation
- big brothers to the TO-220 like the TO-247 for greater power handling
- smaller brothers for compact equipment
- and so on.

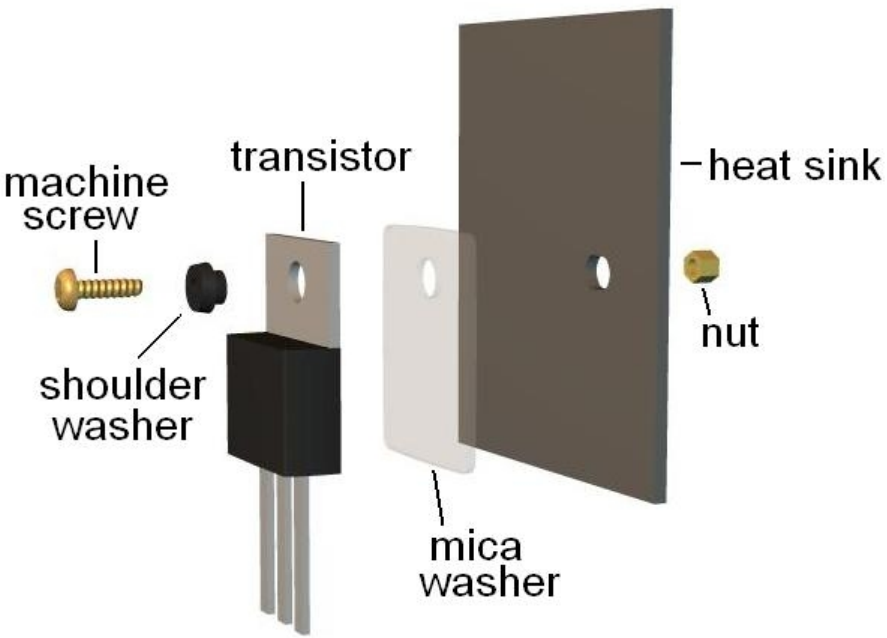
Often, instead of mounting a transistor with a screw, a manufacturer will opt to hold it to the heat sink with a clip.

INSULATED MOUNTINGS

The reason a high-power solid state device has a metal or partly metal case is so it can transfer heat to a heat sink. This presents a problem. The metal part of the case is electrically connected to the collector/drain of a transistor, the anode of an SCR, or one end of a TRIAC, and that part often cannot be grounded. This means, if we mount the device directly to the heat sink, the whole heat sink will have to be insulated from ground, and will be electrically "hot." That's considered poor practice.

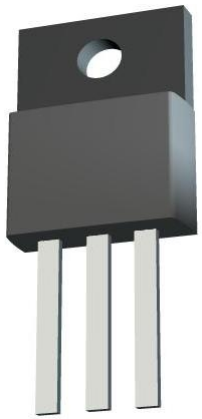
What we need is a way to connect the device to the heat sink thermally but not electrically. Back in EM1, we noted that most electrical insulators are also heat insulators. However, **mica** is so hard, and such a good electrical insulator, that a very thin sheet of it can be used to electrically isolate the device from the heat sink. Even though mica isn't a good heat conductor, the sheet is so thin that heat gets through anyway.

The standard way of mounting a power device to a heat sink is to use a mica sheet specially made with a hole in it to accommodate a machine screw. A plastic shoulder washer is also required to keep the screw itself from contacting the device or sink. Heat transfer is enhanced by putting a dab of heat sink "grease" on the device and the heat sink.



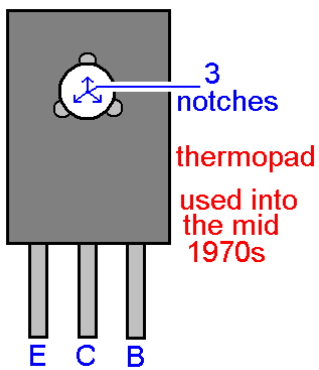
A service tech needs to keep a tube of heat sink grease on hand to re-grease both sides of the the mica sheet when replacing a bad power rectifier or other high-power device. General-purpose grease will not do a satisfactory job.

Over time, the grease will dry out, which is probably the most common cause of transistor overheating and failure.



Some power devices are made with the metal tab totally enclosed in plastic. You still need to use grease, but you don't need the insulators. When a power transistor has an all-plastic case, be aware that there must be a certain amount of “give” in the mounting system, usually supplied by a cup washer. Don't just use a machine screw and nut. If you do, as the transistor case expands with heat, it will crack against the unyielding screw and nut.

Occasionally you'll find a power transistor with leads reversed . Before the TO-220 lead arrangement was standardized, Motorola made “Thermopad” transistors with, left to right, emitter, collector and base leads. These are all plastic, with a metal surface on the bottom – no protruding tab. They can be identified by the little notches molded into the hole in the plastic.



SCR's and Triacs were occasionally made in TO-3 cases, but today almost all of them are found in TO-220 and similar cases.

It's worth finding some junk equipment and pulling it apart to see how it's assembled. Especially note how manufacturers have found clever ways to mount power transistors with leads inserted in PC boards and tabs screwed to heat sinks.

HIGH-POWER CASES

Very high-power devices have special cases. One type is the “hockey puck” case used for rectifiers, SCR's and Triacs. This is a fat disc with a machined face on both sides, to permit sandwiching between two heat sinks. Hockey puck cases are mostly found in older, very high power and high voltage equipment.

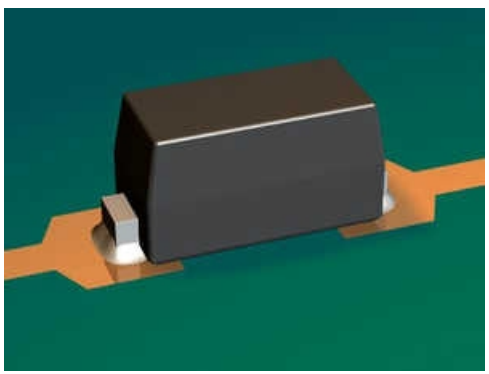


IGBTs often have large cases that are screwed to a heatsink, with screw terminals on top for the electrical connections. Only well-trained, experienced techs should service equipment with these devices in it. You're working with lightning here.

No high-power device that I know of comes with enough built-in heat sinking to dissipate the power the manufacturer says it can handle. For example, in free air, a TO-220 device can dissipate at most a couple of watts. The manufacturer assumes it will be attached to a heat sink. It's the equipment designer's job to be sure an adequate heat sink is provided. It's the service tech's job to be sure that, in equipment he/she is working on, the high power devices actually are transferring heat efficiently. If a unit has been in service for a few years, you may want to clean and re-grease the mica washers.

Some manufacturers make heat sink washers that combine fiber and various rubber-like compounds. If you find a piece of equipment with this kind of washer and no grease, it's safe to assume no grease is required.

SURFACE-MOUNT CASES



In most new equipment, smaller parts are soldered onto the surface of the PC board without bothering with holes. The standard surface-mount case for low-power transistors is called, not surprisingly, the **SMT** case.

High-power surface-mount cases tend to be variations on the TO-220 model. The main differences are the solder tab, and leads that are formed for surface mounting instead of insertion in holes. Any tech who is good at soldering can replace high-power SMT devices.

Small-signal devices require special training, a steady hand, a good light source and a microscope. Depending on your career path, you will either get special training and equipment to work on SMT devices – or you won't work on them at all, simply swapping out whole subassemblies when repair is needed.

PART NUMBERS

There often seems to be no rhyme or reason to part numbers. The truth is, there is some rhyme and reason, but not much. In the USA, there is an electronics industry association that assigns standard part numbers to solid-state devices called the **JEDEC Solid State Technology Association**. JEDEC is an acronym for the association's old name, the **Joint Electron Devices Engineering Council**. You won't be quizzed on that in the ETA exam.

JEDEC part numbers identify the number of PN junctions in a device. A diode, which has only one junction, starts with "**1N**" as in 1N4001, a 1-amp, 50V rectifier. A transistor, which has two junctions, would start with "**2N**" as in 2N3055, a widely used power transistor in a TO-3 case. The numbers which follow 1N or 2N were assigned sequentially and have nothing to do with the part's specifications.

Japanese transistors follow a different system, which manufacturers in some other Asian countries have adopted. A transistor number that begins with **2SA** or **2SB** is a PNP transistor. One that begins with **2SC** or **2SD** is an NPN transistor. **2SJ** identifies a P-channel FET and **2SK** identifies an N-channel FET. Often, the "2S" is omitted, so a transistor marked "C1234" would be a 2SC1234.

Finally, there are many manufacturer-specific numbering schemes. For example, a part number beginning with "IR" means that this part was first manufactured by International Rectifier, though many manufacturers now make equivalent parts with the same numbers.

Often, a standard part number will be followed by a letter, as in 2N3055A. The letter indicates a variation in the case or a special electronic characteristic. For most purposes, all variations are interchangeable, but it pays to check.

For example, General Electric used to make transistors with A, B, and C designations that correspond to 60V, 80V and 100V voltage limits. (And pretty red and green cases too!) If you replace a “C” transistor with an “A” transistor it's likely to blow up. And I've already mentioned “R” which means a stud rectifier whose polarity is the reverse of what you'd normally expect. Stay alert.

Many parts also include a date code, with two digits to identify the year and two to identify the week, e.g. “7414” means the 4th week of 1974. The problem here is that part numbers and date codes can be confused. You may have to refer to a circuit diagram to find out what the part numbers mean.

And now, on to chapter 3, where you'll actually learn to *use* these cool parts!

End of Chapter 2

EM3 Chapter 3
TRANSISTOR AMPLIFIERS

Objectives of this chapter:

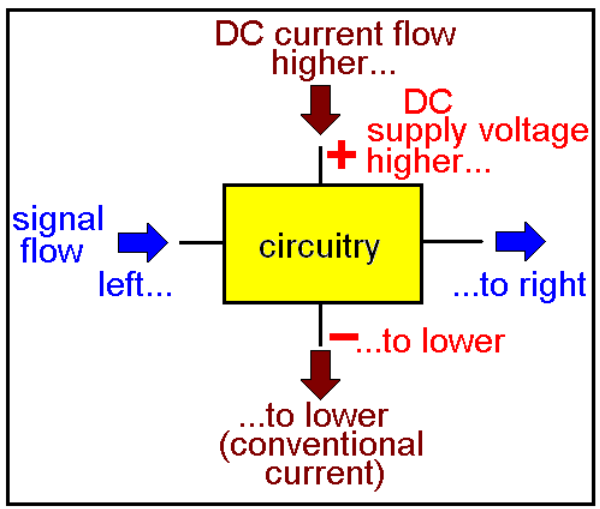
- *describe the three configurations of transistor amplifier circuits*
- *discuss biasing and and why it is done*
- *explain and compare BJT and FET biasing*
- *describe the design of BJT and FET amplifiers in each of the three configurations*
- *discuss degeneration and negative feedback*
- *explain how gain is calculated*
- *discuss class A, B, C and D amplifiers*
- *discuss audio, video, RF, IF amplifiers*
- *discuss the use of transistors as switches*

We've reached a kind of milestone. You studied DC and Ohm's Law in EM1, AC in EM2, and signals and active devices in the first part of EM3. Now we're going to pull the covers off electronic equipment and see how their circuits work. By the end of this chapter, you will be able to take schematic diagrams for some of the less complicated kinds of real-world equipment and trace how the signal flows and understand how each part of the circuit works.

A single transistor is called a **discrete device** to distinguish it from one that combines many transistors in a single package, which is called an **integrated circuit** or **IC**. There is little difference between a discrete transistor and a transistor found in an IC.

SCHEMATIC DIAGRAMS

Everybody has their own way of drawing diagrams. I want to take a moment to discuss some of the things I do, that other people may do differently.

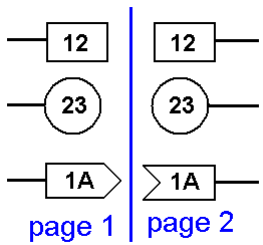


As far as possible, I draw higher DC voltage points higher on the page, with (conventional) DC current flowing from higher on the page to lower, like water going downhill, and I draw signal flowing from left to right. I've found this habit extremely helpful in understanding circuit operation.

You will often run across equipment diagrams so badly laid out that it's almost impossible to trace signal flow. I find that taking a pencil and paper and redrawing those diagrams my own way helps a lot. By the time I'm done, I've learned how the circuit operates. Something you should try when you run into this problem.

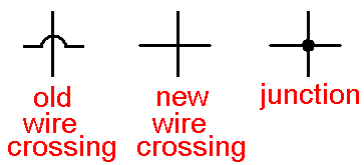
In some of the schematic diagrams for this chapter, I'm going to use the battery symbol to indicate a DC power source, whether or not that supply would be a battery in the real world. In others, I'm just going to use an arrow to indicate an attachment to a power supply or ground point. This is commonly done in actual equipment schematics.

In complicated equipment where diagrams fill several pages, numbered rectangles or other shapes indicate connection points between sections. None of the diagrams in this chapter are that complicated, but when you get into the field, you'll see it a lot.



As in previous chapters, I'm often going to use the word **rail**, which means a wire that distributes DC power to all the circuitry in a unit or connects all the circuitry to ground.

I am once again revealing my old fogey-ness by the way I handle wires that cross each other. It used to be that wires crossing without joining were indicated by little half-circles. Nowadays, people have gotten lazy; the lines just cross. If they are supposed to be connected at the crossing, that's indicated by a dot.



Sometimes, when a diagram is displayed on a low-resolution monitor or photocopied several times, it's hard to tell whether there's a dot. I'm going to use the old method to avoid ambiguity. I'm also, for the most part, going to use the positive current convention.

CHAPTER 3 ORGANIZATION

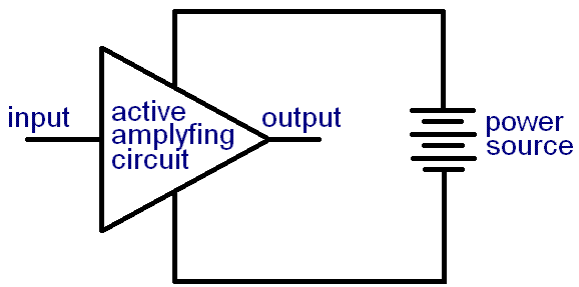
Another long chapter. I suggest you skim through it once before you really dig into it. Chapter organization:

1. **Amplifier discussion**
2. **BipolarTransistor amplifier operation**
3. **The most common BJT amplifier configuration**
4. **Other BJT amplifier configurations**
5. **JFET amplifier operation**
6. **MOSFET amplifier operation**
7. **FET amplifier configurations and how they compare to BJT amplifiers**
8. **Amplifier classes**
9. **Complementary devices**
10. **Transistors used as switches**
11. **Amplifier types and applications**

AMPLIFIERS

Before we talk about transistors, we need to talk about amplifiers in general. An amplifier is a circuit that increases magnitude of an electronic signal, ideally without changing it in any other way. It consists of an **active amplifying circuit**, a **power source** (almost always DC), an **input** to accept a signal to be amplified and an **output** to pass the amplified signal on to the next part of the system.

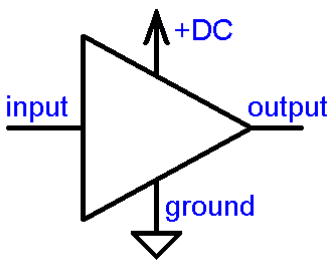
The general symbol for an amplifier is a triangle, usually drawn with the signal input on the left and the signal output on the right. The same symbol is used in both schematic and block diagrams. As you probably know, **amp** has become a widely used abbreviation for "amplifier." You can usually tell from how the word is being used whether it means "amplifier" or "Ampere."



Note that, in the diagram above, the actual voltages at the battery connections are not labeled. In a particular circuit, the positive connection may be grounded, the negative may be grounded, the ground may be halfway between, or there may be no ground. The important thing is that there must be a DC voltage potential across the amplifier to make current flow. And, in most cases, the input and output voltages should never go higher than the positive DC voltage, even for a moment, or lower than the negative DC voltage. This is especially important with opamps, which we'll study in the next chapter.

Note also that the diagram at left obeys the rule: higher DC voltage higher on the page, signal flowing left to right.

In block diagrams, the circuit that supplies power to the amplifier is not shown, since the diagram is only intended to show signal flow. In a schematic diagram, the power supply circuit is usually shown, but is usually moved to a different part of the diagram to avoid visual clutter. The amplifier connection to the DC supply rail is indicated by an arrow, and a ground connection is indicated by the usual ground symbol.



The amplifier symbol is oriented so the input is represented by one of the broad sides of the triangle, while the output is the opposite point. This may seem strange, as the cross-section of the triangle gets smaller from input to output – you'd think it would get larger to indicate amplification.

The reason it's drawn this way is that an amplifier often has several inputs, so the symbol has to make room for them, but there's usually only one output.

As noted in chapter 2, the process of amplifying is called **amplification**, which is a mouthful – try saying it five times fast – so we usually substitute the word **gain**. The amount of gain is stated either as a multiple, e.g. 10X or 40X, or in decibels. An amplifier can boost current, voltage or both. Of course, since power is voltage times current, an amplifier that boosts either voltage or current automatically boosts power. Amplification is often abbreviated $\mathbf{A_V}$, $\mathbf{A_I}$ or $\mathbf{A_P}$ depending on what it's designed to amplify.

The heart of an amplifier, the part that actually does the signal boosting, is always an active device or circuit. You don't need an active device to reduce the strength of a signal; all you have to do is throw some of it away via a passive voltage divider. You recall that this is called **attenuation** or **loss**. But to amplify a signal, an external source of power is required.

This is a very important thing to understand:

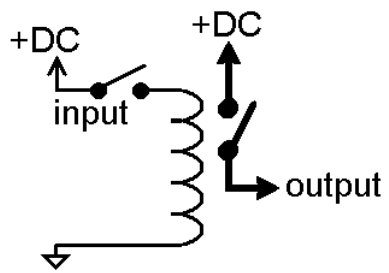
The power in an amplifier's output signal does not come from the input signal, it comes from an external power source.

We can't just magically multiply the signal; that would be a something-for-nothing machine and it doesn't exist in nature. If it did, we could just feed a 9V battery through a series of amplifiers until we had enough power to run the whole world. We wouldn't have to build power plants and worry about the fuel they consume and the pollution they create. Nice dream, but there ain't no such thing.

If we want the output signal to have more voltage, current or power than the input signal, we have to get it from somewhere else. For example, let's say we have a 1mW signal and we want to boost it to 10mW. That means we have to pull in at least 9mW from a power source, probably more because no amplifier is 100% efficient.

In most cases, the input signal makes no contribution at all to output power. Think of a car: you control the speed by pushing your right foot down. The pressure you apply with your foot is not part of the force that moves the car – you, personally, are not pushing the car along in any way. You're just supplying a command to the engine. Likewise, the input signal supplies only the waveform pattern. The amp draws enough power from the DC supply to create a larger copy of this pattern and feed it to the output.

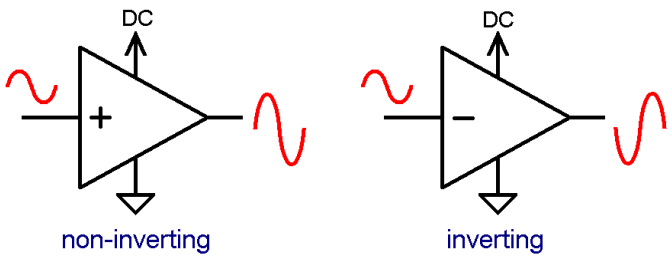
A relay can be thought of as a crude kind of amplifier. A wimpy set of contacts closes to pass a small current through the coil, which closes a brawny set of contacts that pass a large current. Not one bit of the current through the brawny contacts comes from the current in the coil; it comes from somewhere else – a battery or power plant.



A bipolar transistor does about the same thing: it's a current valve, the main current path is through collector and emitter, the valve is opened by a current through base and emitter, and collector current can be a lot higher than base current. The difference, of course, is that a transistor doesn't have to be all “on” or all “off.” And it can operate millions of times faster than a mechanical relay.

INVERTING AND NON-INVERTING AMPLIFIERS

In the course of amplifying a signal, some amplifiers reverse its polarity: when the input goes in the positive direction, the output goes in the negative direction. These are called i**nverting** amplifiers. In other amplifiers, the output waveform is the same polarity as the input. These are called **non-inverting** amplifiers. Which is more desirable, or whether it even matters, depends on the application. In a schematic or block diagram, a “+” at the input indicates a non-inverting amp and a “-” indicates an inverting amp. If there's no polarity indication, we're usually dealing with an application where polarity doesn't matter.



In this tutorial I'm going to ignore vacuum tubes. Tubes make good amps, but they're rare nowadays. Most techs will only ever deal with transistor amps.

BIPOLAR JUNCTION TRANSISTOR AMPLIFIERS

You already know how BJTs work; now it's time to learn how they're used in actual amplifier circuits. Is it really necessary for a technician to know that? Yes, because you need to be able to look at a circuit designed by somebody else and understand how it works. Otherwise, you won't know how to use the voltages you measure to decide which part is bad. Sometimes they smoke, but not always.

Our amplifier discussion will begin with bipolar junction transistors, BJT's. After we've got the basics down, it will be easy to apply them to JFET and MOSFET amplifiers. If it's been a while since you read chapter 2, you may want to go back and review it.

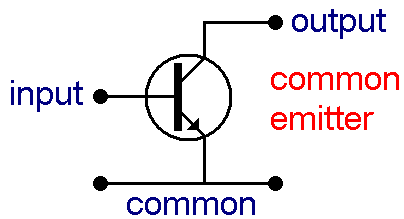
Many people like to think of a transistor as two circuits – an input circuit and an output circuit, just like a relay. The most important concept to bring with you from Chapter 2 is that a transistor is an *electrically-operated current valve*.

Most of this chapter deals with NPN transistors, simply because there's no reason to do a separate study of PNP transistors. They work the same way as NPN circuits if you reverse the polarities. However, later in the chapter, I'll show you some circuits that use NPN and PNP transistors together.

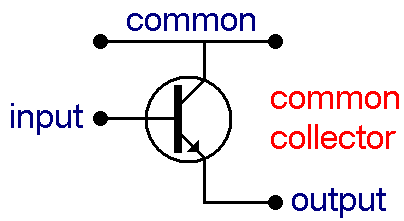
BASIC AMPLIFIER CIRCUITS

If a transistor has separate input and output circuits, but only three leads, that means one lead has to be common to both input and output. There are three ways a transistor can be connected:

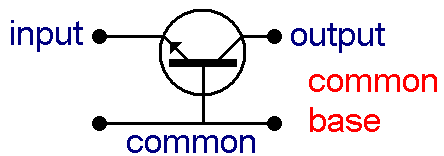
- **Common emitter**, where the input signal is applied to the base, the output signal appears at the collector, and the emitter is connected to the common.



- **Common collector**, where the input signal is applied to the base, the output signal appears at the emitter, and the collector is connected to the common.

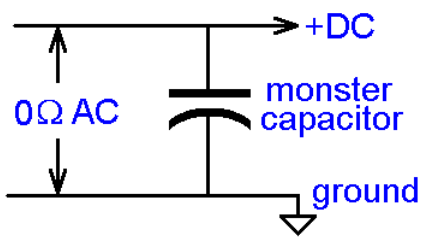


- **Common base**, where the input signal is applied to the emitter, the output signal appears at the collector, and the base is connected to common.



I'm going to use the abbreviations **CE**, **CC**, and **CB** for the three circuit configurations. You won't often see these abbreviations in general usage - “CB,” even among techs, typically means Citizens Band radio – but I refuse to type out those long names again and again. Yeah, I'm lazy.

The common transistor lead is not necessarily at DC ground potential. A transistor needs a DC voltage across it to operate, and in CC and CB circuits, the common points are not at zero volts DC. However, it is possible for any point in a circuit with any DC voltage on it to be at AC ground. In EM2, we talked about **bypass capacitors**. Quick review: a capacitor is an open circuit at DC, but passes AC. The higher the frequency, and/or the bigger the capacitor, the lower its reactance. If we connect a big capacitor between a point in the circuit and ground, the capacitor will essentially short them together at signal frequencies. From an AC standpoint, that point is a ground.

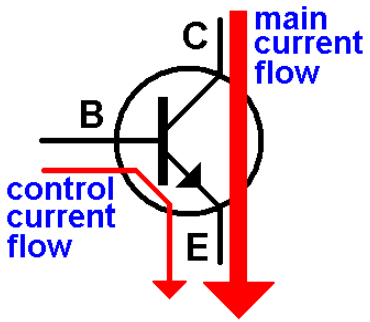


It is customary to AC-ground a DC supply rail in this manner. Some electronic circuitry demands this kind of AC grounding, some doesn't – but it's always good practice to do so. In these study guides, I will always treat DC rails as if the AC impedance between rail and ground is zero, except for specific cases which I'll call to your attention as they arise. It's poor practice to actually use the DC rail as a signal ground, because this theoretical condition is never quite met in the real world. But the AC-grounded rail concept will help as we analyze amplifier circuits.

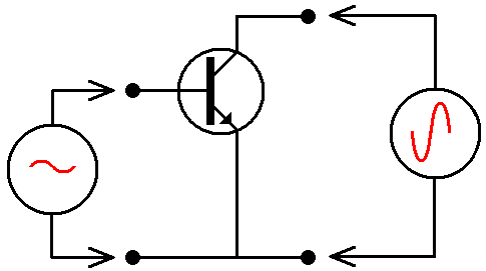
We're going to study CE amps first. CE is by far the most-used configuration. There's a good reason for that: the CC configuration only amplifies current, the CB configuration only amplifies voltage, but the CE amplifies both. A particular transistor will usually give us more overall gain in the CE configuration than in the other two. For that reason, CE can be thought of as the “standard” configuration, and the others as “special-purpose” configurations.

Review from chapter 2: a circuit's ground is the common voltage reference point for all parts of the circuit. The ground's voltage is defined as zero volts, for both AC and DC voltages, and all other AC and DC voltages in the circuit are measured against this reference.

In the CE configuration, the emitter is connected to ground. The controlling current (positive convention) flows into the base, down through the emitter to ground. The much larger controlled current that results flows into the collector, down through the emitter to ground. Thus the emitter is common to both circuits. In a CE amp, “ground” and “common” are the same thing.



The base is the BJT's control terminal, and in an amplifier, the AC signal is applied across base and ground. The collector is the output terminal; the AC signal appears between the collector and ground.



In most cases, the input and output signals will be measured in AC Volts. The exceptions are when significant output power is required, in which case the output signal will be measured in Watts, and when (rarely) we want an output current measured in AC Amps. To start with, we'll be dealing with voltage amplifiers.

For this circuit, the signal common is also the DC voltage ground, which by definition is 0V. The supply voltage source must be more positive than the common to make current flow in the collector. For all the example circuits in this chapter, we'll set the DC source at +10VDC.

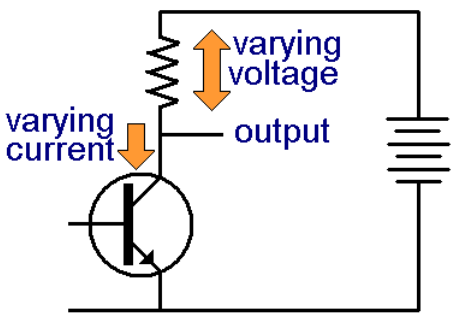
A BASIC 1-TRANSISTOR CE VOLTAGE AMPLIFIER

Amplifier design can be viewed as a process of solving a series of problems. We have to solve the first few to get the amp to work at all. The first step is to clearly state our goal: we want to take a low-voltage AC signal and feed it into the base of our CE amp, and get a much higher-voltage signal out at the collector. Ideally, there should be no change in the signal waveform other than size.

Circuit design is often best approached backwards: start with the signal you want to wind up with at the output, then design backwards through the circuit to get it.

Problem 1: the AC signal at the collector will be a varying current. We want the output to be a varying voltage.

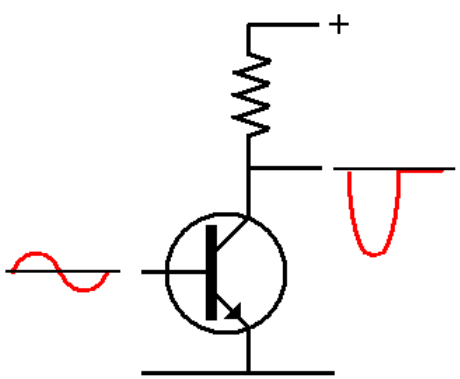
Solution: as mentioned in chapter 2, we derive a varying collector voltage from the varying collector current by putting a resistor between the DC rail and the collector. For current to get to the transistor, it has to flow through this resistor, and that will create a voltage drop. If collector current varies, the voltage drop will vary. This will give us an AC voltage at the collector which we can take as the output signal. Strictly speaking, the output voltage appears between collector and the DC rail, but that's the same as appearing between the collector and common because the DC supply either has a naturally low AC impedance or is bypassed.



You probably don't need to be told that, in this circuit, resistor current equals collector current. They're in series, after all, and you learned way back in EM1 that current is the same everywhere in a series circuit. So, if collector current varies, resistor current varies, meaning the voltage drop varies, so now we have an AC output voltage. Problem 1 solved.

The BJT is normally off. At this point, collector current is zero, so there is no voltage drop across the collector resistor. Voltage at the collector, V_C, will be 10V.

Let's apply a sine wave to the base an see what happens. As the cycle begins, the voltage starts to go positive. When it reaches 0.7V, the base-emitter junction starts to conduct. This causes the collector-emitter circuit to start conducting, which causes a voltage drop across the resistor. So, for the first half-cycle, we get a voltage change at the collector which is like the voltage change at the base, but larger.



Note that the collector pulls the bottom end of the resistor down toward ground, resulting in a half wave whose polarity has been flipped over. **A common-emitter amplifier is an inverting amplifier.** Sometimes that polarity reversal is a problem, but most of the time it isn't. In this case we'll assume it's ok.

Now let's continue along the input sine wave to the zero crossing, and then into the negative alternation. Base current, and thus collector current, is zero at the zero crossing. The current can't go lower than zero, so during the negative alternation, nothing happens. Thus...

Problem 2: the transistor amplifies the positive input alternation, inverting it in the process, but doesn't amplify or even pass the negative alternation. The output waveform isn't anything like the input waveform. Houston, we have a problem.

Solution: our problem is that collector current can't go lower because it's already zero to start with. We need to set the circuit up so collector current is flowing at all times. Then it can increase when the input signal goes in the positive direction and decrease when the input signal goes in the negative direction. If we set things up so collector current never goes to zero during any part of the signal waveform, we'll get the whole cycle at the output.

This is called **biasing** the circuit. The constant current flow, in the absence of an AC signal, is called **bias current** or **quiescent current** – "quiescent" because it's the current that flows when the circuit is quiet, i.e. when there's no AC signal present. When a transistor has been biased, the various DC voltages and currents are collectively referred to as the **operating point**.

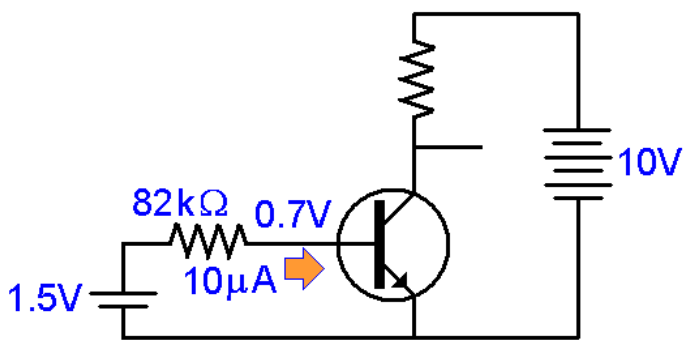
How much collector current do we want? That gets kind of involved. In most cases it's a somewhat arbitrary decision. Just for grins, let's set it at 1mA, a typical value for an amplifier that handles small signals.

Since the base is the control terminal, to make current flow through the collector we start by making a smaller current flow through the base; specifically, the collector current we want divided by the transistor's current gain. In a common-emitter amplifier, the transistor's current gain is identified by the symbol β, the Greek letter **beta**. Assume we have a transistor with a beta of 100, which is a pretty common real-world value. That means, to get a collector current of 1mA, we need a base current of 0.01mA, or 10uA. (Remember, the differences between units, milli, micro, nano, pico are 3 decimal places each.) 10uA is 10^{-5} A.

When a BJT is "on," that is, conducting current, the base-emitter junction is a forward-biased diode. The emitter is grounded at 0V, which means the base is at about +0.7V for a silicon transistor. To make that 10μA of current flow through the base, we need to provide a DC voltage source that's higher than 0.7V. One way of doing this is to provide a separate battery. A typical single-cell carbon-zinc battery is around 1.5 volts, and we already know the base-emitter junction is a forward-biased diode, so we need a resistor between the battery and the base that conducts 10μA as it drops 1.5V – 0.7V = 0.8V. Ohm's Law time:

$$\begin{aligned} R &= V / I \\ &= 8 \times 10^{-1} V / 10^{-5} A \\ &= 8 \times 10^{4} \text{ Ohms} \\ &= 80\text{kOhms.} \end{aligned}$$

The closest standard value is 82k.



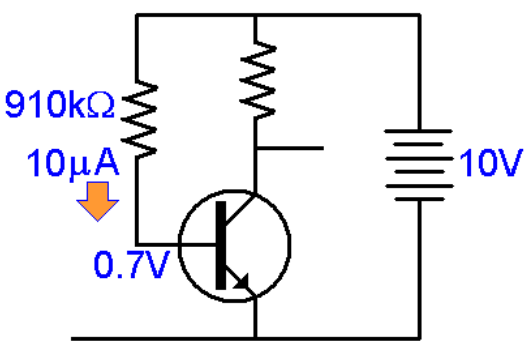
But wait, this is stupid! We've already got a +10V source. Instead of adding another battery, let's supply the base current from the 10V supply. We just have to increase the value of the dropping resistor, because we're now dropping more voltage.

Supply: **10V**
Base voltage: **0.7V**
Difference: **10 – 0.7 = 9.3V**

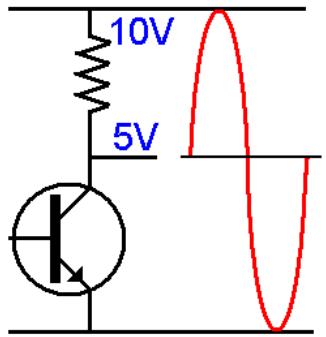
We've already calculated that we need a current of 10^{-5}A. So,

$$9.3V / 10^{-5}A = 9.3 \times 10^{5} \text{ Ohms}$$
$$= 930kOhms$$

The closest standard value is 910k.



We've got one thing left to do: choose the collector resistor. It is customary to set the collector voltage at half the supply voltage. The resistor can't pull the output any farther up than the +10V rail, and the transistor can't pull it any farther down than the ground rail. So, if we set the collector voltage at 5 volts, that allows for a maximum voltage swing in both directions.



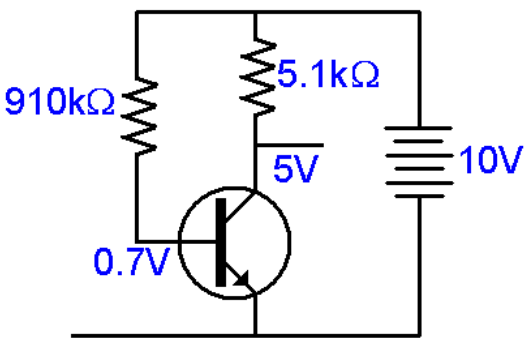
Now, we won't be looking to the transistor to do this. The transistor has no control over its own collector voltage; it only controls current. So we have to look to the collector resistor.

The top end of the resistor is connected to the 10V rail. We want the bottom end to be at 5V, so the resistor needs to drop 5V. We've chosen a 1mA collector current. By now, you should find this as easy as naming a rock musician with a giant ego:

$$R = V / A$$
$$= 5V / 10^{-3}A$$
$$= 5 \times 10^{3} \text{ Ohms}$$
$$= 5kOhms$$

The closest standard value is 5.1k.

SIGNAL IN AND OUT



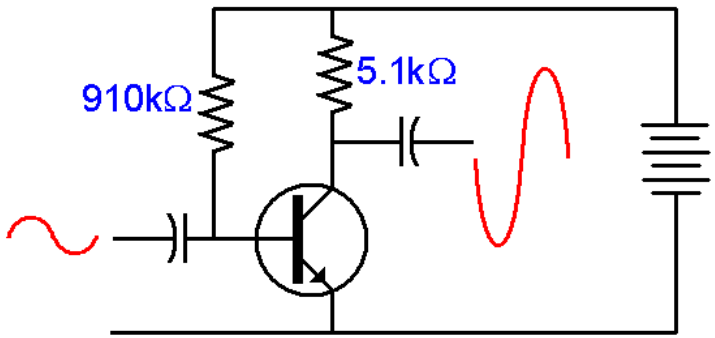
Now that we have set up a bias point, we can apply an AC signal to the base to vary it. Problem 2 solved.

Problem 3: the signal source is referenced to common, and probably at 0VDC, while the base is at +0.7VDC. We need to connect the AC signal to the base, but we have to keep the DC voltages separate.

Solution: enter our old friend the capacitor. Recall that a capacitor passes AC but blocks DC. If we insert a capacitor between the signal source and the base, it will couple the AC but block the DC. You'll recall that, when a capacitor is used this way, it's called either a **coupling** or **blocking capacitor**.

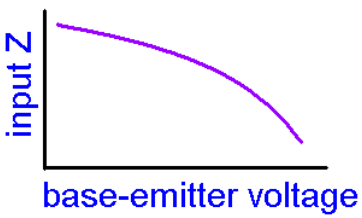
The same problem occurs at the output; we want to get the AC signal out, but we don't want the 5 volts of DC that comes with it. So we insert another capacitor. Problem 3 solved.

The final circuit is shown above right. If we put it together, will it work? Yes. Lots of equipment has circuits exactly like this inside. But it's pretty crude. We can do a lot better.

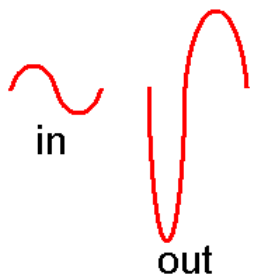


BJT AMPLIFIER NONLINEARITY

Problem 4: the base-emitter junction is a diode. That means, during negative input signal alternations, the input impedance will go in the direction of an open circuit, and during positive alternations, it will go in the direction of a short. In other words, the input impedance and thus the load on the signal source is nonlinear.



The voltage-vs.-current curve of a diode is also nonlinear. This means the signal voltage to base current conversion will be nonlinear, so the waveform at the collector output will be distorted.



Solution A: if the input signal is a small enough fraction of the 0.7V junction voltage, nonlinearity won't be a problem. Up to about 30mV p-p at the input, the distortion is insignificant unless we're dealing with a signal, such as music, where very high waveform accuracy is required. So, with small signals, this problem doesn't crop up.

Solution B: sometimes we don't care about distortion. The telephone system has always had lousy sound quality, and it works just fine for its intended purpose. In the 2-way radio world, anything under 5% distortion is considered okay.

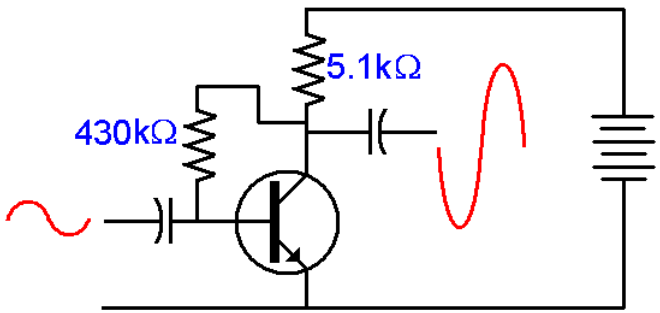
Solution C: distortion can be reduced by adding resistors into the circuit. This also reduces gain, which we will probably have to make up elsewhere in the circuit. It turns out that we often want to add resistors anyway, to stabilize the operating point. So I'm going to put off further discussion of waveform distortion for a bit, until we've talked about stabilization.

COMPENSATING FOR BETA VARIATIONS

Problem 5: As noted in chapter 2, beta varies a lot between transistors, even ones with the same part number. We've been designing for a beta of 100. Suppose our transistor has a beta of only 50, meaning that a 10ua base current will result in a collector current of only 0.5mA. Now, instead of the collector sitting at 5V, it will sit at 7.5 volts, and we'll only be able to get half as much signal voltage out before the collector voltage hits the 10V rail. Worse, suppose beta is 200. Now collector current will be 2mA. A 2mA current flowing through a 5kΩ resistor will result in a 10V drop. This means the quiescent collector voltage will be zero, the transistor will be saturated, and only the negative signal alternations will come through.

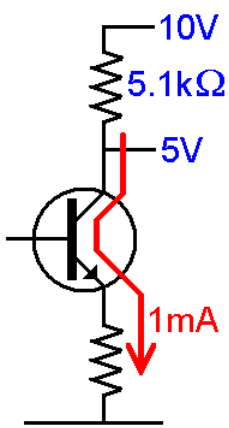
Solution A: measure individual transistors until we find one with a beta of about 100. There are some circuits where this kind of sorting is the only option.

Solution B: change the circuit so it automatically compensates for beta variations. The simplest way to do this is to connect the base dropping resistor to the collector instead of the 10V supply. Now we're dropping 5V – 0.7V = 4.3V and the dropping resistor will be 430k.



If the beta is 100, all is well. If it's greater, collector current will increase and collector voltage will fall. This will reduce the voltage drop across the base resistor, reducing base current and partially correcting the error. The opposite will happen if the beta is lower than 100. Either way, the collector voltage will diverge from the 5V target, but not as badly as before.

There's another way to stabilize the operating point that begins with putting a resistor in the emitter circuit. Before, the emitter was always at ground, but now it's at a voltage above ground. That voltage varies when emitter current varies, due to plain old stupid-simple Ohm's law at work.



First we calculate the emitter resistor. We need to know emitter current, which is the sum of collector current plus base current, $\mathbf{I_E = I_C + I_B}$.

If beta = 100, I_B is only 1% of I_C, so it hardly counts. **Rule of thumb: for all practical purposes, $\mathbf{I_C \sim I_E}$** except when we're dealing with a really crummy low-gain transistor.

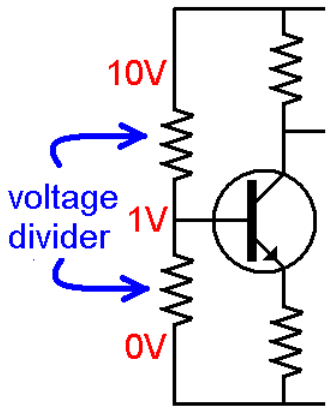
Choice of emitter voltage is somewhat arbitrary. The more voltage we drop across the emitter resistor, the more stable the circuit will be. However, the transistor cannot pull its collector down farther than its emitter. That being the case, whatever voltage we put on the emitter is subtracted from maximum collector voltage swing, reducing maximum signal output. So we don't want the emitter voltage to be too high.

For the heck of it, let's set the voltage drop across the emitter resistor at 0.3 volts; that will put a convenient 1V on the base. Setting that voltage higher would improve stability – you'll understand why shortly – but maybe in this application perfect stability doesn't matter. Since I_C and therefore I_E are 1mA, the emitter resistor will be:

$0.3V / 1mA = 300\Omega$, which is a standard value.

Next: when the emitter was grounded, we couldn't control collector current by controlling base voltage. However, now we can. Any variation in collector (and therefore emitter) current will cause the voltage across the emitter resistor to change. If we set a stable 1V at the base, we should get a stable 0.3V at the emitter, which translates to a stable 1mA of current.

So, we firmly establish base voltage by providing a voltage divider which passes much more current from +DC to ground than it passes from +DC through the base. This makes variations in base current largely irrelevant.



If the transistor's beta is 100, we already know we need 10uA for the base. Let's set current through the lower voltage divider resistor at 4 times this value, or 40uA.

$$1V / 4 \times 10^{-5}A = \tfrac{1}{4} \times 10^{5}\Omega = 25k\Omega.$$

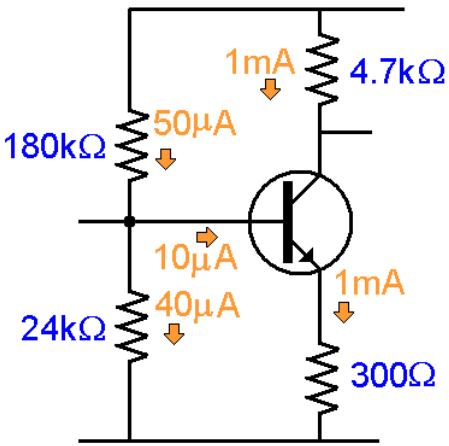
24k is a standard value.

The upper resistor will have to pass this 40uA current plus the 10uA base current, for a total of 50uA. (Kirchoff: total current entering a node equals total current leaving.) Voltage drop is the 10V supply minus the 1V at the base, or 9V. So:

$$R = 9/5 \times 10^{-5}$$
$$= 9/5 \times 10^{5}$$
$$= 1.8 \times 10^{5}$$
$$= 180k\Omega$$ which is a standard value.

If beta is different from 100, collector/emitter current will want to change. This sets a chain of events in motion:

1. Voltage across the emitter resistor starts to change.
2. Thus the voltage between emitter and ground starts to change.
3. Since base voltage is an unvarying 0.7V above the emitter voltage, base-to-emitter voltage will start to change.
4. A small change in the voltage across the base-emitter diode, say to 0.68V or 0.72V, will cause a large change in base current; that's the way diodes are.
5. This will cause a change in emitter current that will counteract the initial change in collector-emitter current.



Thus the departure from the operating point will immediately be almost completely corrected.

DEGENERATION

Putting a resistor in the emitter circuit causes a decrease in voltage gain. This is called **degeneration**.

An interesting characteristic of a BJT amplifier with a resistor in the emitter is that the voltage gain of the circuit is approximately equal to the ratio of collector to emitter resistors, at least for gains of 30 or less. After all, the current through both is about the same, so any change in current will be about the same; thus the respective voltage drops will be in proportion to the values of the two resistors. In our example circuit, the ratio is 4,700 to 300, resulting in a gain of about 14, minus maybe 10% because the transistor itself has small base and emitter resistances we haven't accounted for.

I hope you're alert right now, because we just tripped over one of those concepts that will be fundamentally important in everything that follows. If we start with no degeneration (emitter resistor = 0) and then gradually raise the resistance of the emitter resistor, we reach a point where control of the circuit gain passes from the transistor to the parts around it. Let me repeat what I said above: voltage gain of this circuit is approximately equal to the ratio of collector to emitter resistors. In other words, **gain is controlled by the parts that surround the active device. The device's own gain becomes irrelevant.**



As Martha Stewart would say, that's a good thing, since transistor gain varies a lot from one to the next. A resistor, on the other hand, is accurate to 5% or better, its resistance is stable, and it is linear. This is how we get uniform performance from nonuniform transistors. At a gain of 20 – that is, a circuit where the emitter resistor is 5% of the collector resistor – *any* NPN BJT will work.

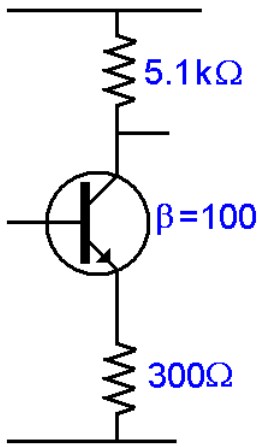
There is a condition: the transistor must be capable of a lot more gain than we want from the final circuit. That way the resistors are the dominant factor. And, of course, if we're dealing with high frequencies, we must choose a transistor that has adequate gain at those frequencies.

If those nice, linear resistors are now in control of the gain, we would expect that problem 4, nonlinearity, would also be eased. And that's what happens. The input impedance at the base is now a lot less like a diode and a lot more like a resistor. So, we use this circuit when a grounded-emitter circuit would cause too much distortion.

Degeneration always reduces amplifier gain. Often, the gain is so low we have to send the signal through a second amplifier. But transistors and resistors cost pennies these days, so, big deal.

INPUT IMPEDANCE

When the emitter is grounded, it's a little tricky to figure out the base input impedance, especially since it varies. But when there's a resistor in the emitter circuit, it's easy: the impedance at the base is roughly equal to the emitter resistor times beta.

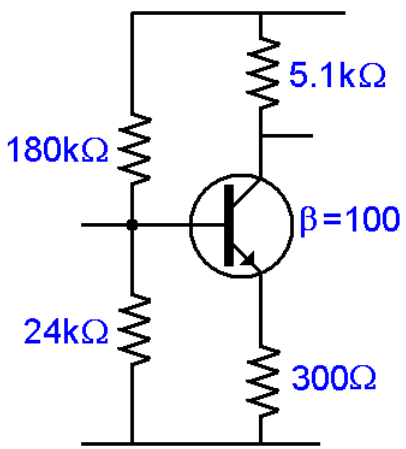


$Z_{IN} = 300\Omega \times 100$
$= 30k\Omega$

It works like this: since the base-emitter voltage is a fairly constant 0.7V, we can say that the AC signal at the emitter "follows" the AC signal at the base, with a constant 0.7V DC offset. In other words, the two AC voltages are equal. Now, the varying AC voltage drop across the emitter resistor causes a varying resistor current, which becomes a varying emitter current. And what causes a varying emitter current? A varying base current! But we know that the base current is much smaller than the emitter current. With a beta of 100, the base current is 1/100th of the emitter current. That means any *change* in base current will be 1/100th of the resulting change in emitter current.

This is the same as saying that the impedance, looking into the base, will be much higher than the emitter resistor – approximately the emitter resistor times beta. In our example, 300Ω x 100 = 30,000 Ohms or 30kΩ. (I say "approximately" because the transistor has some resistance by itself, so the actual input impedance will be a bit higher.)

At this point, an engineering professor would object that I've left some minor terms out of the equation, that I've been sloppy, that the analysis above is only a first approximation. Darn right, I'd reply. Since transistor beta varies by maybe three to one from one transistor to the next – thus input impedance in this circuit varies by three to one – what would be gained by sharpening the pencil? And I'd remind him that trying to get answers more precise than the data you're working with is futile.



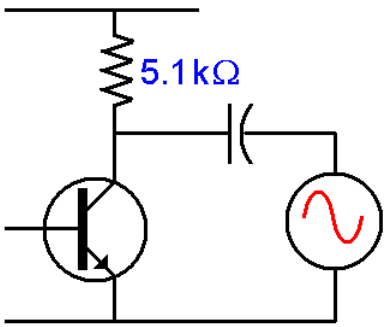
$$Z_{IN} = \frac{1}{30k\Omega} + \frac{1}{24k\Omega} + \frac{1}{180k\Omega}$$

Now, the input impedance we've just calculated is for the transistor base only. Let's calculate the input impedance of the whole amplifier, which is essentially three parallel resistors connected between the base and ground. One is the transistor's input impedance, 30kΩ. The second is the resistor from base to ground, 24kΩ.

The third is the 180kΩ resistor from the 10V supply to the base. We assume the 10V supply is bypassed to ground. So, the input impedance is the parallel resistance of 30k, 24k and 180k. Take out your calculator and go to work. No, I'm not going to tell you, I want you to figure it out.

OUTPUT IMPEDANCE

I've said that collector current isn't affected much by collector voltage. That means, if we connect a signal source to the collector and force a signal backwards into the amplifier's output, making the collector voltage go up and down, collector current will not change. In other words, the transistor does not load the signal source, meaning it's an AC open circuit.

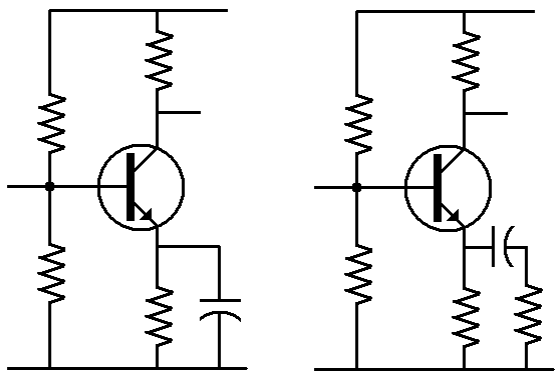


But we're looking at a 5.1kΩ resistor at the same time. Since the transistor is connected between the signal source and ground, and the resistor is connected between the signal source and the DC rail, which is the same as ground from an AC standpoint, the output impedance is the parallel combination of transistor and 5.1kΩ resistor. The transistor's output impedance is essentially infinite, so the output impedance is simply 5.1kΩ. Or maybe 5% lower, around 4.9kΩ, because with real-world transistors, current does vary a little with voltage.

This holds true when the bias is supplied by a single rail-to-base resistor, or in the voltage-divider bias circuit we've been looking at. As we will see, there are other circuit conditions that drastically change the output (and input) impedance.

COUNTERACTING DEGENERATION

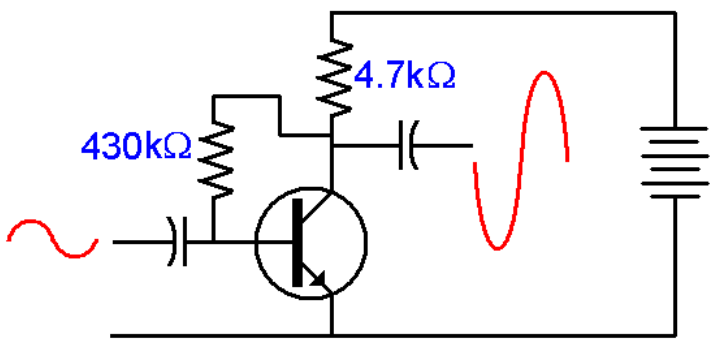
We can achieve the DC stability of a voltage-divider biased circuit with an emitter resistor, but the gain of a grounded-emitter circuit, by connecting a capacitor across the emitter resistor. As you probably guessed, this is called **bypassing** the resistor. The capacitor has to be large enough to have only a few ohms of X_C at signal frequencies.



Since we're essentially grounding the emitter at AC frequencies, we've lost the distortion-reducing benefit of degeneration. We can get something between grounded-emitter gain and degenerated gain by bypassing only part of the emitter resistor, or putting a resistance in series with the bypass capacitor.

NEGATIVE FEEDBACK

Let's look at the collector-to-base resistor arrangement again. As we've seen, the reason for connecting the resistor this way is to stabilize the operating point. Any change in collector voltage will show up as a change in current at the base. This change will act to counter the deviation from the desired operating point. In other words, it's a corrective.



This is called **negative feedback**. The public has adopted this expression, taking it to mean criticism or disapproval.

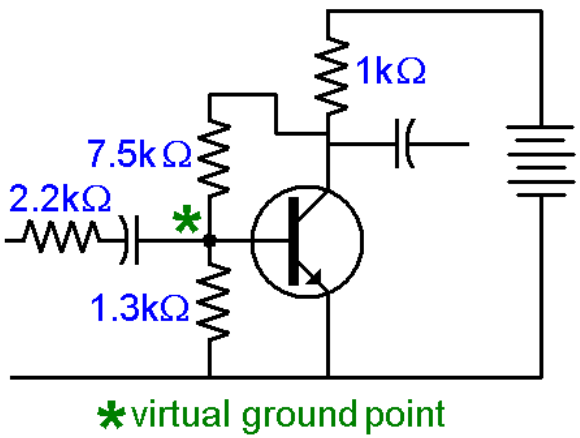
But what it originally meant, and what it still means to technical people, is taking a voltage or current sample at the output and routing it back to the input in such a way as to counteract error.

Negative feedback also works to correct AC errors – in other words, it reduces waveform distortions, just as degeneration does. And, like degeneration, negative feedback reduces gain, which is why it's sometimes called **degenerative feedback**. For feedback to be "negative," its polarity must be the opposite to of the input signal, so it partially cancels the input signal. A CE amplifier inverts the polarity, so that requirement is automatically taken care of.

Negative feedback can increase or reduce input impedance, depending on the circuit. In the collector-to-base resistor configuration at left, it tends to lower it. An instantaneous change in base current caused by the input signal is countered by an opposite change in current through the collector-base resistor. The net effect is to create an artificially low input impedance.

We can also combine the collector-to-emitter and voltage divider techniques as in the circuit at on the next page. In this case, we may not need the emitter resistor because stabilization is achieved via collector-to-base feedback.

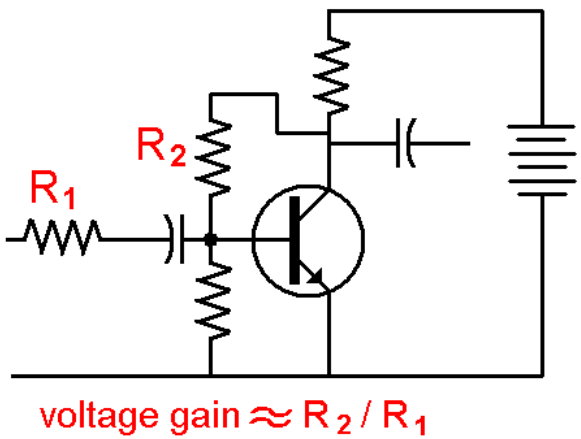
Note that I've lowered the two voltage divider resistors. This gives us a nice, stable operating point that hardly depends at all on beta.



Lowering these resistors has another effect: reducing the impedance at the base to almost zero. With such a low feedback resistance, 7.5kΩ, any change in base voltage is almost completely canceled by a change in current through that resistor. This is called a **virtual ground** input because it isn't a real ground, but it acts like one.

What good is an amp that shorts out the input signal? Well, remember that the almost-zero impedance is created by a varying collector voltage, which creates a varying current through the 7.5kΩ resistor. In other words, we still have an output at the collector. The difference is, the base effectively senses AC current, not voltage. This is sometimes called a **current amplifier**.

We can still use this as a voltage amplifier but, since the amplifier senses current, voltage gain depends on the impedance of the signal source, which can get tricky. However, if we put a resistor between the base and the signal source – 2.2kΩ in the circuit at left – and we assume the signal source has a much lower impedance than that resistor, so it doesn't affect our calculations, and if we don't try for a gain above 20X, then the circuit gain is about equal to the ratio of the two resistors:



$$A_V = R_2/R_1$$

Since the circuit gain is controlled mostly by the resistors, not the transistor – and, since resistors are linear – we have a stable, low-distortion amp. Like degeneration, when a lot of negative feedback is used, the resistors take control of circuit behavior, and transistor beta variations and nonlinearities become almost irrelevant.

MAXIMUM VOLTAGE AND CURRENT GAIN

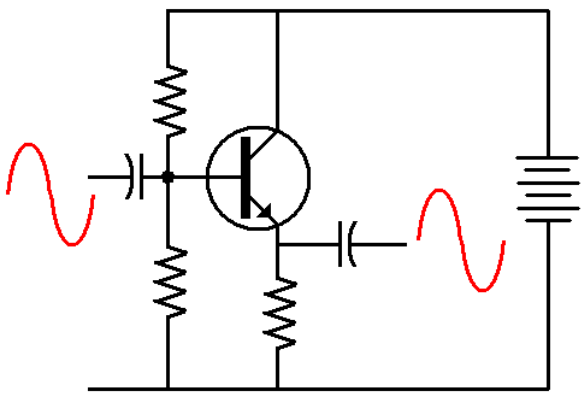
There is a limit to a one-transistor amplifier's voltage gain. We've been pretending the base-emitter voltage is a constant 0.7V, but in fact it varies a little with current. After all, we couldn't really apply an AC voltage across the base and ground of a grounded-emitter amplifier if the base-emitter diode didn't have a little “give” to it. This means a small AC input voltage can only cause so much current change in the base, thus only so much current change in the collector. Bottom line: the voltage gain of a real-world common-emitter amplifier is limited to about 50X in most cases. And, the higher the signal voltage, the more waveform distortion the amplifier will create.

OUTPUT IMPEDANCE WITH NEGATIVE FEEDBACK

I'm not going to get into calculating output impedance in a negative-feedback circuit, because it gets kind of involved, something an engineer has to know but a tech doesn't. However, the effect of negative feedback is to lower the output impedance. Before, collector voltage had no effect on collector current. Now it does, because there's feedback from the collector to the base. This feedback will tend to counteract any voltage change at the collector caused by, for example, an external load. An engineer would use this to design an amp whose output voltage is not much affected by changes in load impedance (up to a point).

Now that you've got a good grasp of CE amplifiers, it's time to look at the other two configurations, and then we'll get into some really clever stuff with multiple-transistor amps.

THE COMMON COLLECTOR CONFIGURATION



In the common collector or CC configuration, the collector is connected directly to the DC supply. The input signal is applied to the base and the output signal is taken from the emitter; thus the collector is the common element. As we've already seen, since there's a relatively constant 0.7V drop from base to emitter, the emitter voltage will follow the base voltage, with a constant 0.7VDC offset. For this reason, a common collector amplifier is often called an **emitter follower**. The signal polarity at the output is the same as at the input, thus this is a **non-inverting** amplifier.

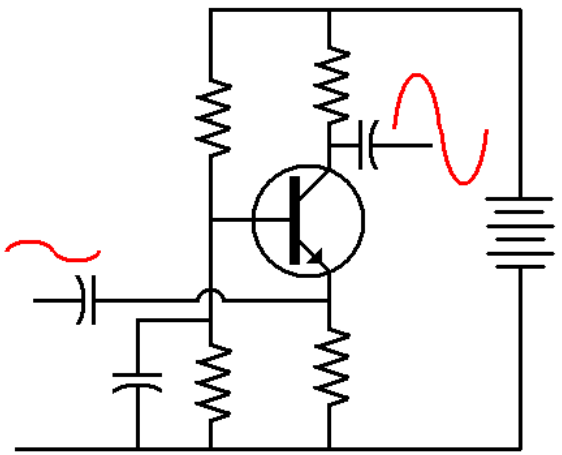
Saying the output voltage follows the input voltage is the same as saying that the AC voltage gain of the amplifier is **unity**, that is, 1X. (In practice, there's a loss of a few percent.) Thus, the emitter follower only amplifies current. And, of course, since $P = VI$, a current amplifier is also a power amplifier, whether we were aiming for that or not.

The CC amplifier's output impedance is very low: it's the sum of the transistor's base resistance and the signal source impedance, divided by the transistor's current gain, which works out to 50Ω or less in most cases. We use it where we don't need to change the voltage of a signal, but want to drive a low impedance load from a high impedance source. For example, the voltage needed to drive a pair of headphones is about the same as consumer-level line voltage, about 0.3V. However, line impedance is high, and headphone impedance is low, usually 32Ω. So we could send the line signal through a CC amplifier to drive the headphones, and many people do.

Or, instead of a low output impedance, we can take the benefit in the form of a high input impedance at the base, which is the emitter resistor times beta. We calculate it the same way we did for the CE amp, except in this case, since the output comes from the emitter, the emitter resistor is typically dropping half the supply voltage instead of the 0.3V we chose before.

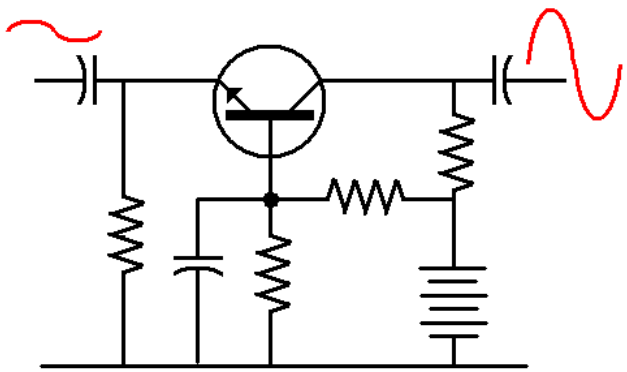
For comparison, let's set a CC circuit up with the same 10V supply and 1mA collector current. If we set the emitter at 5V, the emitter resistor will be 5kΩ. If Beta = 100, that translates to a 500kΩ input impedance. Of course, we have to take into account the shunting effect of the biasing resistors. But the circuit input impedance will still be much higher than with the CE circuit.

THE COMMON BASE CONFIGURATION



In the common base or CB configuration, the emitter is the signal input, the collector is the signal output, and the base, which is common to both circuits, is grounded. Since the emitter must be 0.7V more negative than the base and the collector must be more positive, this suggests that we need to provide two voltage sources. Instead, we put a positive DC voltage on the base, and bypass it with a capacitor.

If a positive-going signal is fed to the emitter, it will tend to decrease the voltage between base and emitter, thus decreasing base current, thus decreasing collector current. When collector current falls, voltage drop across the collector resistor will be reduced, and the output voltage will go in the positive direction. Thus the CB amplifier is **non-inverting**.



Some people draw the CB configuration this way. If you trace it out, you will see it's exactly the same. Go on, do it, you need the practice.

All the AC current in the collector must be supplied by the signal source, so the current gain is unity – a little less, actually, because the signal source must also supply the small AC base current. Thus the current gain, which is called **alpha** (Greek letter α) in the CB configuration, is always just a little less than 1; the CB amplifier only amplifies voltage. Input impedance is usually just a few ohms; output impedance is essentially equal to the collector resistor.

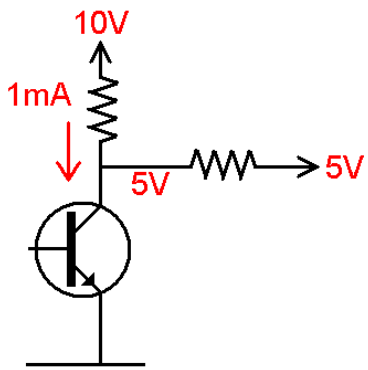
A given transistor will work at higher frequencies in the CB configuration than either of the others. This isn't as important as it once was, now that we have transistors that will work at GHz frequencies in a CE circuit. However, when we get into multiple-transistor amps, we'll find that the common base circuit fills the frequent need for a non-inverting amp with voltage gain.

POWER AMPLIFICATION

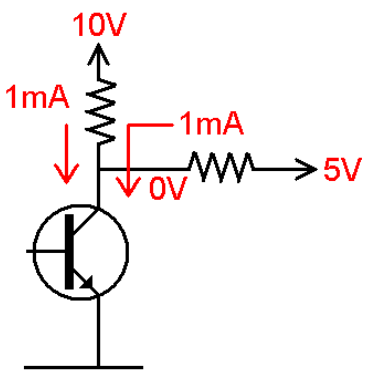
Many people insist that a circuit isn't an amplifier if it doesn't amplify voltage; thus the CC configuration isn't an amplifier. This is splitting hairs, and anyway I disagree. Since power is voltage times current, a circuit that amplifies current also amplifies power, and nobody would claim a circuit that boosts power isn't an amplifier.

When an amp is going to deliver substantial power to a load, a lot of extra complication is introduced. I'll just mention one example here. Let's return to our CE amplifier where the collector sits at 5VDC. The junction between collector and resistor is constantly being pulled up by the resistor and down by the collector.

When collector current increases – when the collector pulls harder – the voltage goes down. When the collector pulls less hard, the voltage goes up.

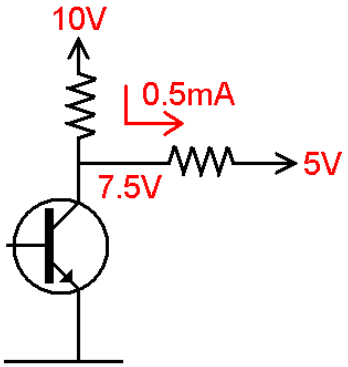


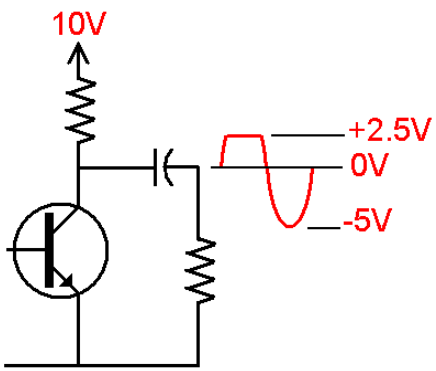
Let's connect a load. Normally, we'd use a coupling capacitor to couple the AC signal to a load resistor whose other end is grounded. But for the moment, I'm going to connect one end of the load resistor directly to the collector, and its other end to a +5V source. That's exactly the same thing, except that it makes analysis easier because there's no DC drop across an output capacitor. In the absence of signal, the collector is also at +5V so there's no voltage across the load resistor and no current flow through it. We'll assume the load resistor is equal to the collector resistor.



Now let's feed in a signal. When collector current increases, the transistor pulls both the collector resistor and the load resistor down. The transistor can pull as hard as necessary to bring the collector down to zero volts.

However, When collector current decreases, the collector resistor pulls up, but can *not* pull as hard as necessary to bring the collector up to 10V. The higher the output voltage goes, the lower the voltage drop across the collector resistor, so the less pull it has. When the voltage reaches 7.5V, collector current is zero, and the collector resistor and load resistor form a voltage divider, splitting the difference between the +10V supply and the +5V point. The voltage can go no higher.



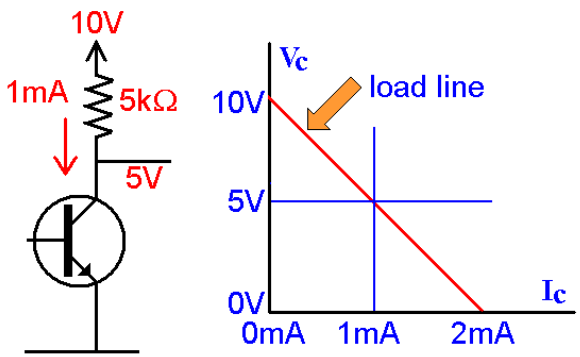


Thus our amplifier can accommodate a 5V negative alternation, but only a 2.5V positive alternation. If we try to get a sinewave with 5V peaks out of it, the positive alternations will be clipped off. In the illustration at left, the load resistor is connected the way you'd expect it to be, via a capacitor, and you can see what happens to the signal.

We can improve the situation by starting with the collector voltage at, say, 3.5V instead of 5V. That puts more voltage across the collector resistor, so it can do a little better job of pulling up. As far as pulling down goes, the transistor won't have any problem pulling the collector down to zero. We still get less than a 10V P-P output, but at least it isn't lopsided.

LOAD LINES

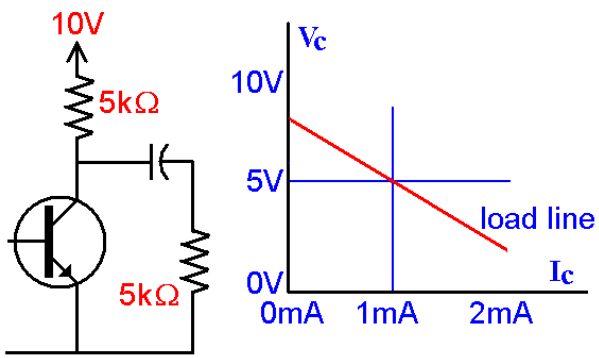
This is a good place to introduce a concept called the **load line**. Let's take the example we've been using – we'll call the collector resistor 5kΩ to make things easy. If we plot collector voltage against current, with current on the X-axis because voltage is a function of current, we'll get a diagonal line between zero current at one extreme and zero voltage at the other. The cross hairs show the quiescent operating point.



The diagonal line is the load line for this circuit; at every point it shows a particular combination of voltage and current.

If current increases, voltage drop increases, so collector voltage decreases. The maximum possible current is 2mA, which drops 10V across the resistor, putting the collector at ground. On the other hand, if current decreases, resistor voltage drop decreases and collector voltage rises. At I_C = 0, there is no voltage drop and V_C = 10V.

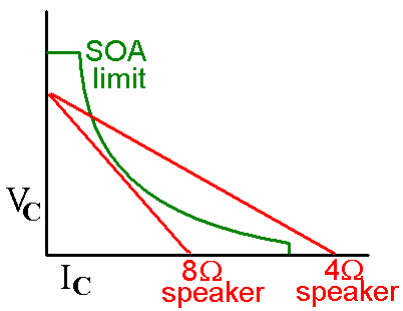
Now let's add the external load. The DC operating point does not change, since the load is connected thru a capacitor and a capacitor is an open circuit at DC. However, when current changes, we only get half as much voltage change as before. Note the smaller load line slope compared to the one at left.



When IC goes to zero, VC only goes to 7.5V, and when IC goes to 2mA, VC only goes to 2.5V. That's because, from an AC standpoint, the two resistors are in parallel, with an equivalent resistance of 2.5kΩ. 1mA of change into 2kΩ of resistance is 2.5V of change.

The load line could go on to 3mA at the right; the transistor is certainly capable of exerting a 3mA pull. But at that point, we've got a lopsided, clipped-off signal, so there's no point.

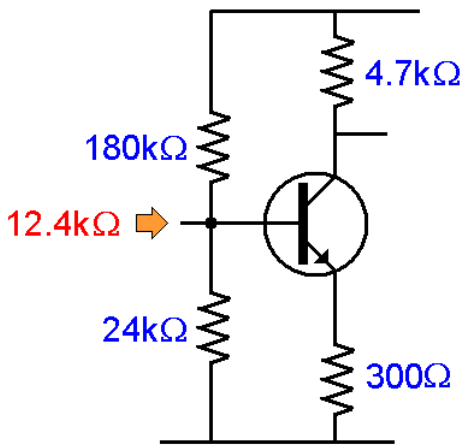
Load line charts are useful for a lot of things. For example, in an audio power amp, we can use one to show the voltage across an output transistor versus current through it for any load. In the chart below, the curved green line represents the safe operating area limit for a 100W transistor. Note that the amp is safe with an 8-ohm speaker (or would be, if the speaker were purely resistive, which is a whole 'nother can of worms) but a 4-ohm speaker will probably blow it up.



We're not going to go into these issues any more deeply at this point. The reason for bringing them up is so you'll be aware that the load affects circuit operation, and so you'll understand why engineers do some of the things you will see them doing.

RAISING INPUT IMPEDANCE WITH A BOOTSTRAP

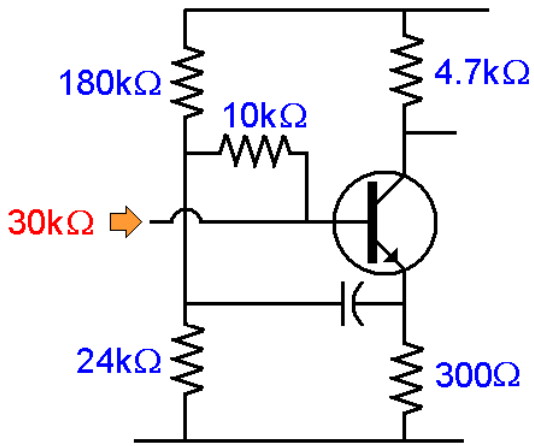
Here's a neat trick. There's an old joke about a guy who wants to fly. So, he reaches down, grabs onto his bootstraps (that will tell you how old a joke it is) and pulls himself up into the air. Okay, it's absurd, that's the whole point, it's a joke. Nevertheless, it has become a standard metaphor for people lifting themselves up in the world by their own effort.



There's an electronic equivalent, and it isn't a joke. In the voltage-divider biasing circuit at left, the transistor accounts for less than half the load on the input; the voltage divider resistors account for most of it. The more firmly the voltage divider controls the voltage – meaning the smaller fraction of total current goes into the base – the more the voltage divider loads the signal source.

When we're dealing with a high-impedance signal source that we don't want to load down, it's aggravating to think that most of the current supplied by the source flows uselessly into the voltage divider resistors instead of the transistor.

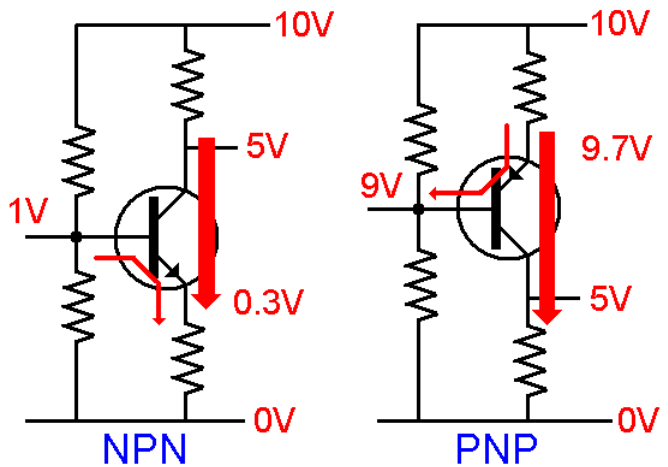
The bootstrap circuit is a way of getting around this problem. It takes advantage of the fact that the emitter voltage follows the base voltage.



The bootstrap circuit at right is a CE amplifier. Note, however, that the biasing voltage divider is tied to the emitter through a coupling capacitor. From the standpoint of DC biasing, this capacitor doesn't exist (it's an open circuit). However, at AC frequencies, from the standpoint of feedback *only*, this is a CC circuit. After all, even though we're taking the output from the collector, that doesn't change the fact that the emitter voltage follows the base voltage like a CC amp. And, due to that coupling capacitor, the biasing voltage divider also follows the base voltage. AC voltage at the base = AC voltage at the emitter = AC voltage at the voltage divider junction.

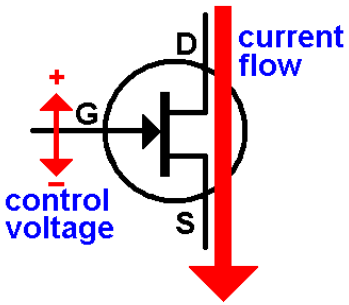
Now look at the resistor between the voltage divider and the base. That's only a 10k resistor; you'd think it would load the input signal rather badly. But because both ends are moving together, the AC voltage across it is about zero, so there's no AC current flowing through it. Thus from the standpoint of the signal, this resistor is almost an open circuit. The transistor's emitter pulls its own bias voltage divider up and down, bootstrap style. The only load the signal sees is the input impedance of the transistor itself.

PNP amplifiers



PNP amps are just like NPN ones but with reversed polarities. Above is a circuit with voltages and conventional current flows marked, so you can see what I'm talking about. Study this carefully; a lot of the equipment you'll be working on contains both types of transistor.

FET AMPLIFIERS

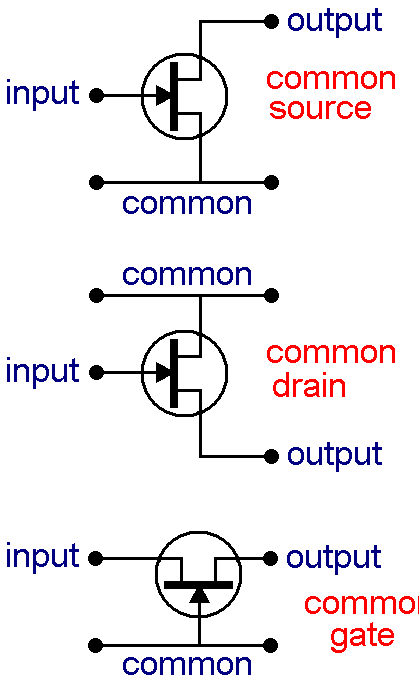


Single-transistor FET amplifiers are like their BJT counterparts, with some differences. A bit of review of the material in chapter 2: first, instead of emitter, base and collector, the connections are called **source**, **gate** and **drain**. Second, a FET is a voltage-controlled device, as opposed to a BJT which is a current-controlled device. That is, current flowing through the device from drain to source (conventional current) depends on the voltage at the control terminal rather than the current through it.

A FET amplifier has the same three configuration choices as a BJT amplifier, though the names are different because the names of the terminals are different. As shown to the right, the **Common Source** or **CS** amplifier corresponds to the BJT CE amp. The **Common Drain** or **CD** amp corresponds to the BJT CC amp. The **Common Gate** or **CG** amp corresponds to the BJT CB amp.

The CG circuit has little reason to exist and is seldom used, which means we can pretty much assume the gate is always the control connection and signal input. The current-flow path is from the drain through the source.

As noted in chapter 2, the gate does not conduct current, except for a very small leakage; in other words, at DC, the input is an open circuit. The AC input impedance is not infinite due to capacitance, which may be a serious problem at high frequencies. But for low frequency amplification and switching, you could hardly ask for a better device: a flea-powered input that hardly loads the signal at all.

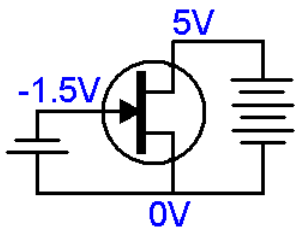


FETs, like BJTs, come in complimentary polarities. We'll be studying **N-channel** FET amplifiers, in which the source-drain current channel is N-doped silicon, and which are the nearest equivalent to NPN transistors. **P-channel** FETs are the same, except the voltage polarities are reversed.

JFET BIASING

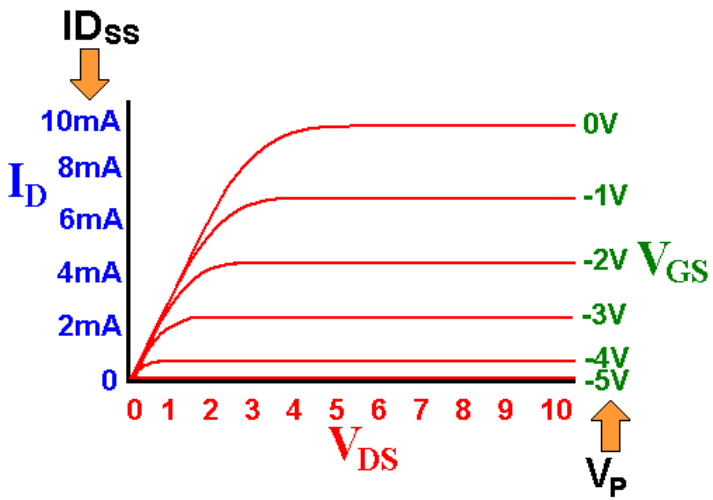
As we saw in chapter 2, FETs that are normally on, and require a reverse voltage at the gate to turn them off, are **depletion-mode** FETs. A JFET is always a depletion mode device.

It is normally “on” and, in the case of an N-channel JFET, its gate voltage must be more negative than the source voltage to turn it off.

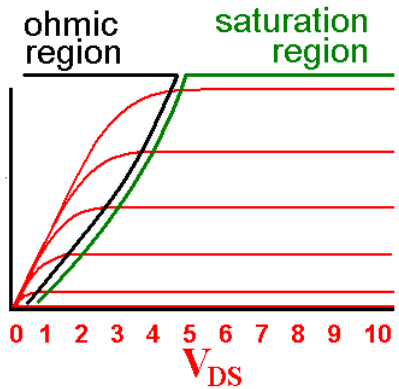


If we plot a typical JFET's behavior, it will look like the chart below. We'll need to talk about this a bit because it's a complicated chart. Current and drain voltage have been flipped from the load line chart, so drain-to-gate voltage is on the X-axis and drain current is on the Y-axis.

The two most important numbers for setting a JFET amplifier's operating point are **ID_{SS}**, the drain current when gate to source voltage is zero, and **V_P**, the gate-to-source voltage when drain current is zero (pinch off). For this particular JFET, ID_{SS} is 10mA and V_P is -5V.



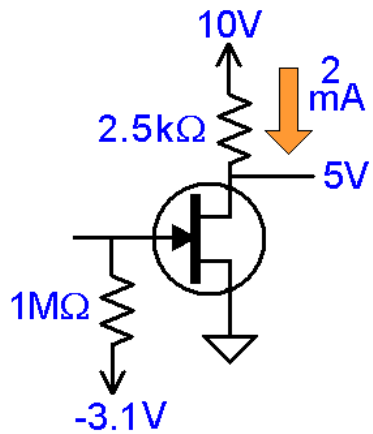
The chart can be divided into two sections. In the **saturation region**, V_{DS} essentially doesn't matter; once drain voltage gets high enough to be beyond the curved sections, a further increase has little effect on current. In this way a JFET is like a BJT. In the **ohmic region**, V_{DS} matters. A change in VDS results in a change in current. In this way a JFET is like a voltage-controlled resistor.

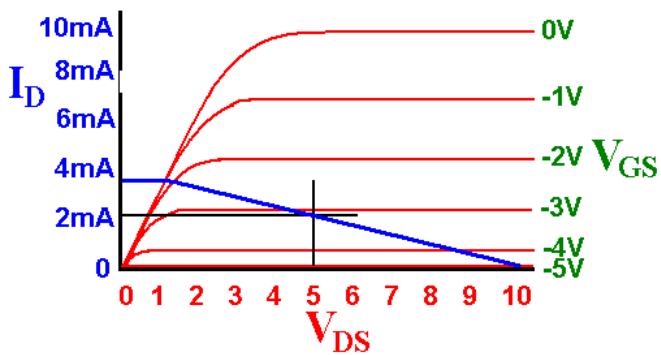


THE SATURATION REGION

Let's look at the right part of the graph first. Assuming V_{DS} is more than a few volts, I_D is controlled almost completely by V_{GS}. This is the saturation region of the chart, but “saturation” doesn't imply the same thing as it does with a BJT – the drain and source aren't at the same voltage. “Saturation” means a physical condition of the charge carriers within the silicon. Note that the V_{GS} / I_D relationship is not linear. A JFET amp without degeneration, like a BJT amp, distorts the signal, with distortion increasing as signal level increases.

Suppose we have decided, based on signal levels and so on, that we want I_D to be 2mA. We have a 10V power supply, so the maximum voltage will be 10V at I_D = 0mA. Out of habit, we'd probably put the drain at 5V. That would imply a 2.5kΩ resistor between the drain and supply. The gate resistor can be very high because the gate draws no current. We take the chart published by the FET's manufacturer and draw cross hairs that indicate the operating point, telling us that if we want 2mA of drain current, we need to set the gate at -3.1V.





Let's figure out the voltage gain. If we pull the gate down to -5V, which is the pinchoff voltage, that's a change of -1.9V. The drain will go to 10V, which is a change of +5V.

So gain is:

$$5 / 1.9 = 2.6$$ or thereabouts.

Huh? Only about an 8dB voltage gain? That stinks!

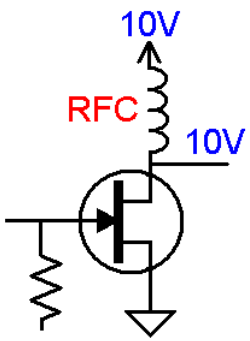
Well, we could do better with a higher supply voltage, which would let us use a higher resistor at the drain. That would increase the gain. And if we didn't really need 2mA of drain current, and we could cut it back to, say, 1mA, we could use a higher resistor yet and that would increase the gain some more. But the fact is, a JFET just isn't a high-voltage-gain device; 20X is about the limit.

On the other hand, it has massive current (and thus power) gain. In theory, the gate's resistance is infinite; in reality, it might be 300MΩ, 3 x 10^8Ω. That puts its current and power gain in the neighborhood of a million – for *one single transistor*. In most cases, the resistor attached to the gate is the only significant load on the signal source.

Even before we did all that biasing work, we could have made a good guess about circuit gain. The spec sheet for this FET lists transconductance in the saturation region as 1mS. That's 1 milli-Siemens, 10^{-3} Siemens, 10^{-3} mho. That means, for every volt of change at the gate, there will be 1mA change in drain current.

With a 2.5k drain resistor, that works out to 2.5V of change. So, if we were talking about an AC signal, 1V in, 2.5V out, or a gain of 2.5X. That's pretty close to the 2.6 we figured out by graphing.

If we're amplifying high frequencies, there's a trick we can use to increase voltage gain, and this works in a BJT circuit too. We replace the drain or collector resistor with a **radio-frequency choke** or **RFC**. Its DC resistance is low, but its AC reactance will be quite high at the frequencies we're amplifying. This will give us a voltage gain of 100X or more if lightly loaded.

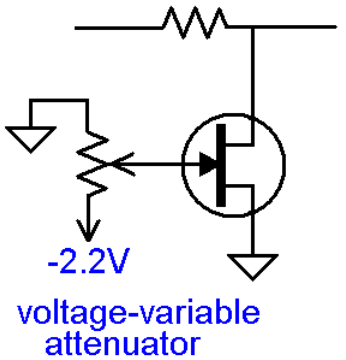


It also puts 10VDC on the drain, due to the coil's low resistance. Thus the JFET can pull down nearly 10V, and flyback will take it up to 20V. A 20V p-p output from a 10V supply! Cool, yes?

THE OHMIC REGION

A JFET will work in the ohmic region; the difference is that changes in V_{DS} affect I_D and gain is lower. I'm not going to describe biasing and operation in the ohmic region because the FET doesn't work well there and most circuit designers just stay away from it.

However, when there is no DC across the FET, it can work as a resistor whose resistance is varied by changing V_{GS} – in other words, a voltage-variable rheostat. This has long been used in such things as variable-gain amplifiers. The signal has to be kept under about 100mV to avoid distortion. In the illustration at right, a pot is used to vary V_{GS}, which doesn't make much sense, it's a pot varying a pot. But the FET's resistance can also be controlled by various kinds of sensing circuits, or maybe by a computer. One such use is to automatically compensate for signal level variations. A sample is taken from the signal, rectified, filtered and fed to the gate. This creates a voltage divider whose attenuation increases whenever the signal increases.



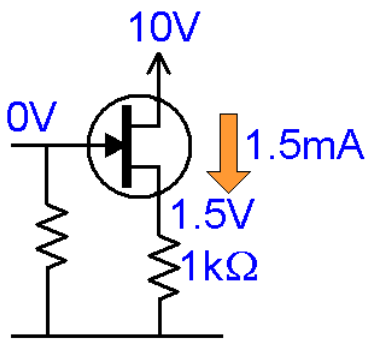
SELF BIASING

Up to now, we've used a positive rail to supply channel current and a negative rail to bias the gate. When a piece of equipment already has positive and negative power supplies, this works fine.

However, most circuits have only a positive supply. The need to provide an additional negative supply just to bias a FET is an expense we would like to avoid.

Instead of biasing the gate negative, we can bias the source positive. The easiest way to do this is to simply put a resistor between source and ground. Current flowing through the channel pulls the source voltage up, which is the same as pulling the gate voltage down. This is called **self biasing** or **automatic biasing** and it's a trick engineers invented back in the vacuum tube days.

In the diagram at right, we're using a JFET that draws 1.5mA when its gate-source voltage is -1.5V. So, if we put a 1kΩ resistor in the source circuit, it will drop 1.5V and the gate-source voltage will be correct.



The source resistor does the same kind of operating-point stabilizing as the emitter resistor in a BJT amp. And that's lucky, because the gate voltage vs. current behavior of JFETs varies as much as the beta of BJTs. For example, a J211, an inexpensive, common JFET that's been around for decades, may have a pinch-off voltage of anywhere from -2.5V to -4.5V. If we don't want to sort through them to find one we like, we absolutely have to design with lots of degeneration to compensate for variation among transistors. And that means even lower voltage gain, unless we bypass the source resistor.

JFETs were used a lot in radio receivers at one time, due to the limitations of other kinds of devices. However, improvements in those other devices have now relegated the JFET to a niche role. The other place JFETs are used by the gazillions today is as input devices in the integrated circuits that we'll study in following chapters.

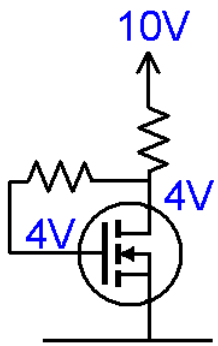
MOSFET BIASING

As you know from chapter 2, MOSFETs can be either depletion mode or enhancement mode. A depletion mode FET is normally "on" and the gate is brought negative (with respect to the source) to turn it off. In other words, in normal operation, a depletion-mode MOSFET is a lot like a JFET. An enhancement mode FET is normally "off" and the gate is brought positive to turn it on. Most MOSFETs are enhancement mode, in part because that eliminates the need for two power supplies.

From here on out, you may assume that when we talk about a MOSFET we're talking about an enhancement mode device.

An enhancement MOSFET is biased by setting the gate at a positive voltage that results in the desired channel current. This isn't a slam dunk. In most MOSFETs, the gate voltage difference between fully off and fully on is fairly small. Moreover, it won't be the same from one device to the next. This is one of the main reasons MOSFETs haven't gained a lot of popularity in linear circuits. On the other hand, they are the ideal active component for switching circuits where they're either all-on or all-off.

The simplest way to bias an enhancement-mode MOSFET is to connect a resistor between drain and gate. Since the gate is an open circuit, voltage drop across the resistor will be approximately zero, even if it's a very high-value resistor like 10 megohms. The MOSFET will bias itself to whatever gate voltage is needed to establish current flow in the drain-source circuit. If we want a higher voltage at the drain, we can bias the gate with a feedback voltage divider.



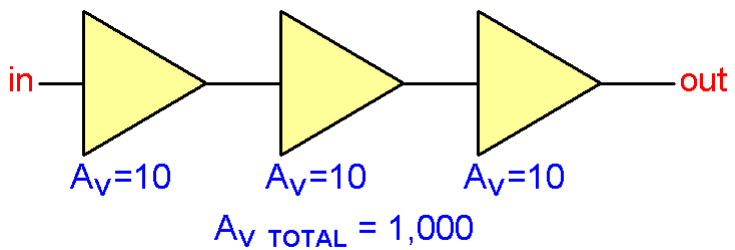
MOSFETs, like JFETs, vary in their gate-source voltage characteristics, so we have to provide for that. We can use source degeneration and voltage divider bias as we did with BJTs. The advantage with MOSFETs is that the divider resistors can be very high-value, so loading on the signal source is minimal.

More precise and controllable MOSFET biasing involves feedback around several transistors. Which brings us to...

MULTIPLE-TRANSISTOR AMPLIFIERS

If one transistor doesn't supply as much gain as we need, we can make amplifiers with more than one transistor. There are so many ways to do this that I can only show you a sampling of the most common configurations.

The most basic multiple-transistor amplifiers are just a series of single-transistor amplifiers. Each single-transistor amplifier is called a **stage** of amplification.



In theory, there is no limit to the number of stages and thus no limit to how much a signal can be amplified. In practice, if a signal is weak enough, it will be buried below the noise floor, and then there is no point in amplifying it because the noise is amplified at the same time.

We're going to begin by studying different ways of making an amplifier with two common-emitter stages. To avoid clumsy language, we're going to adopt the convention used in schematic diagrams: a transistor is identified by the letter "Q," not to be confused with "Q" of a tuned circuit. The signal is amplified first by Q_1, then passed on to Q_2 which amplifies it further.

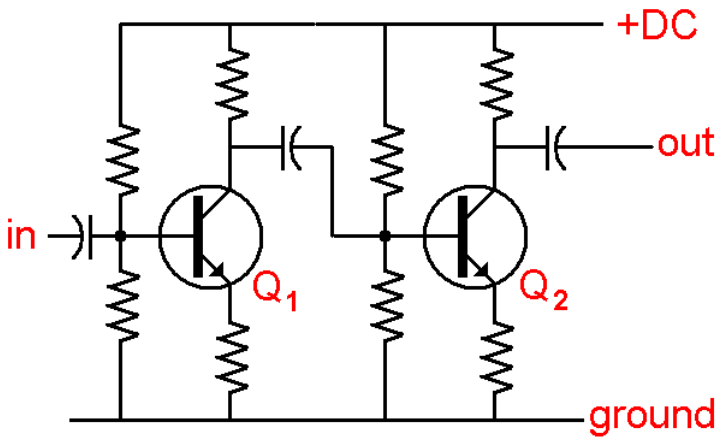
COUPLING

When we have multiple amplifier stages, the issue of coupling arises – how to get signal out of one amplifier and into the next. The voltage-divider biased amp we've looked at has 1VDC at the input and 5VDC at the output, so obviously we can't just connect two such amps in series.

Capacitor Coupling

The easiest way to connect two amplifiers is to put a coupling capacitor between them. This isn't a terribly complicated process:

1. Find the total resistance at that point in the circuit. If two CE amps are coupled together, R_T will be the output resistance of the first amp plus the input resistance of the second.



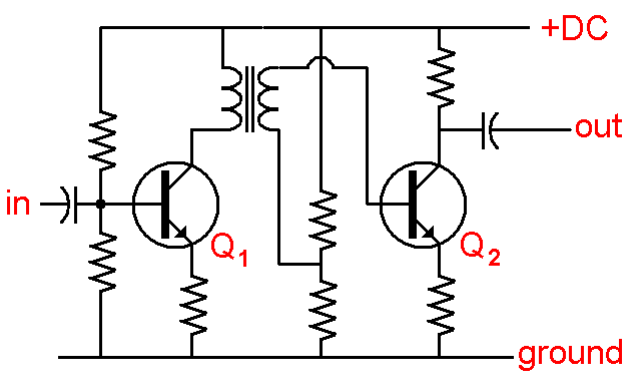
2. Calculate a capacitor value that will pass the whole range of frequencies you want the amp to handle. This means X_C must be equal to or lower than the R_T we calculated in step 1, down to the lowest frequency you're interested in. If you want very little loss, make the capacitor 10x bigger than calculated.
3. If the resulting value is large enough that you have to use an electrolytic cap, be sure it's connected into the circuit with proper polarity.

The circuit above shows a practical way to couple a 2-stage amplifier using input, output and **interstage** coupling capacitors.

In a circuit like this, we should probably choose capacitors according to the 10X rule, because the low-frequency attenuation will be the sum of three capacitors' worth of rolloff.

Transformer Coupling

Another way of coupling amplifier stages is through the use of a transformer. The primary is connected into the collector of Q_1; the secondary is connected to the base of Q_2. Q_2's biasing voltage divider is connected to the side of the secondary opposite from the base. Since the secondary's DC resistance is quite low, this is the same as connecting the divider at the base, but without the signal loading effect.



One advantage of coupling with transformers is that the signal's voltage/current ratio can be changed in the process, and this can increase gain. Back in the 1920s, when tubes had low gain and high cost, transformer coupling was used to maximize gain.

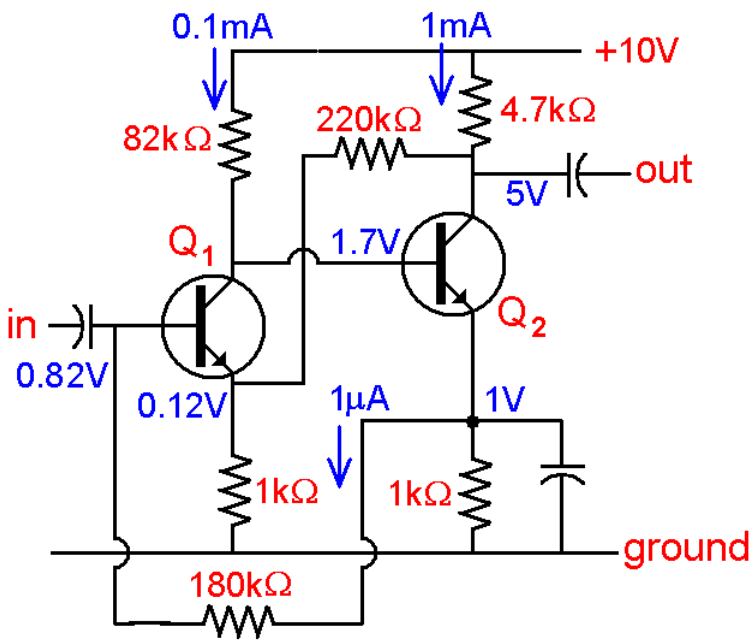
Nowadays, transistors are so cheap that stage gains don't have to be maximized. However, at radio frequencies, transformers are cheap too, and there are additional reasons for using them. For one thing, RF filters often take the form of tuned transformers. Transformer coupling has always been the standard coupling method in radio receivers where tubes or discrete transistors are used.

Direct Coupling

The coupling problem can be avoided entirely by designing the stages so the DC output of the first stage is the same voltage as the input of the second. I'm going to take you through several examples, and you'll want to fasten your seatbelts because this is heavy going. At this point in your study, it may be more than you can handle, but you should keep coming back to this until you've got it. A tech needs to be able to pinpoint a malfunction, and in a direct coupled circuit, you have to know what you're doing, because everything affects everything else.

Direct Coupled Example no. 1

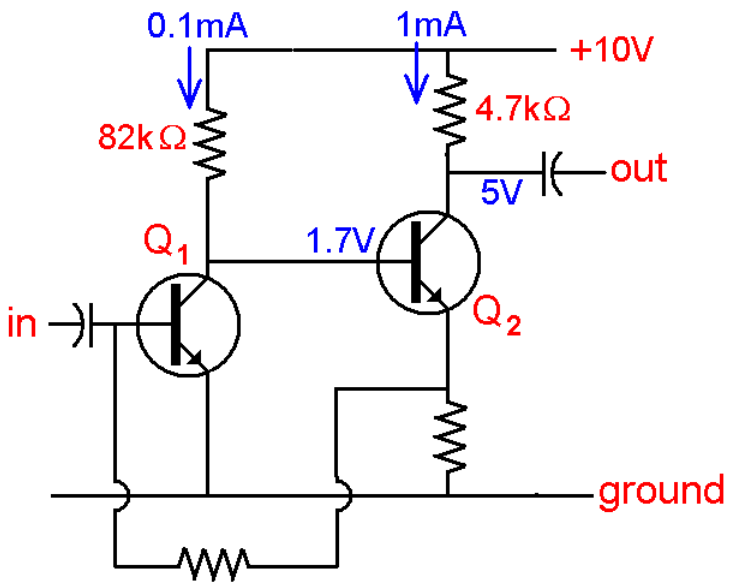
The circuit below is widely used to amplify tiny signals. Voltage gain can easily be 1,000X or more. The collector of Q_1 is connected directly to the base of Q_2, meaning the two stages are coupled at DC as well as AC. Let's look at how it operates.



Hmm, your eyes just glazed over, didn't they? Okay, take deep, calming breaths. It'll all make sense. Promise.

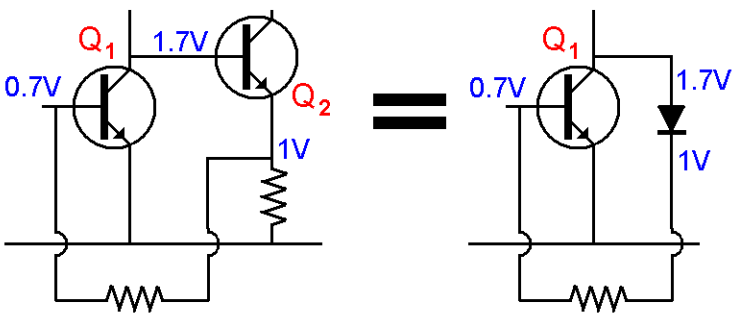
One good way of analyzing a complex circuit is to become temporarily "blind" to its complexity, and look only at the most basic characteristics of its operation. The diagram below is the same circuit, stripped down to its simplest form.

The first thing to notice is that the collector voltage of Q_1 is nowhere near half the DC supply voltage. That means it won't be able to produce much output voltage swing. But then, it doesn't have to, since it's just feeding the input of Q_2.



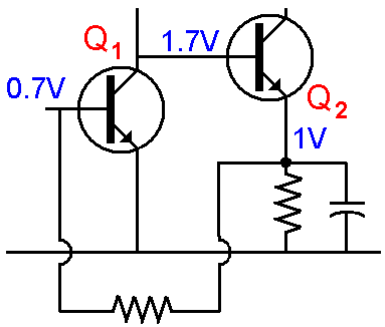
There's a reason it's designed this way. Q_2 is the output transistor; we want the collector voltage to be able to swing up and down as far as possible. But, as noted earlier, it can't swing any further down than its emitter voltage, so we want to keep the emitter voltage low. The base voltage is, of course, only 0.7V higher than the emitter voltage – and that's the reason Q_1's collector voltage has to be low.

Let's look at how the operating point is established. There's a resistor between the emitter of Q_2 and the base of Q_1. Since the emitter of Q_1 is grounded, we know the base is about 0.7V. We assume there's a small drop across the resistor feeding Q_1's base, so the emitter of Q_2 will be around 1V.



If you remember that the base-emitter junction of Q_2 is a diode with a 0.7V drop, you'll figure this out before I tell you: from a DC standpoint, this is equivalent to a virtual-ground amp, except that it has a voltage-dropping diode in the feedback circuit.

In a circuit like this, where feedback is applied around more than one stage, we call the feedback circuit a **feedback loop.** Since there's very little voltage across the bias feedback resistor, any change in operating point will cause a major change in current through this resistor. In other words, we have a feedback loop with very effective error correction. The operating point of the circuit is likely to be pretty stable.

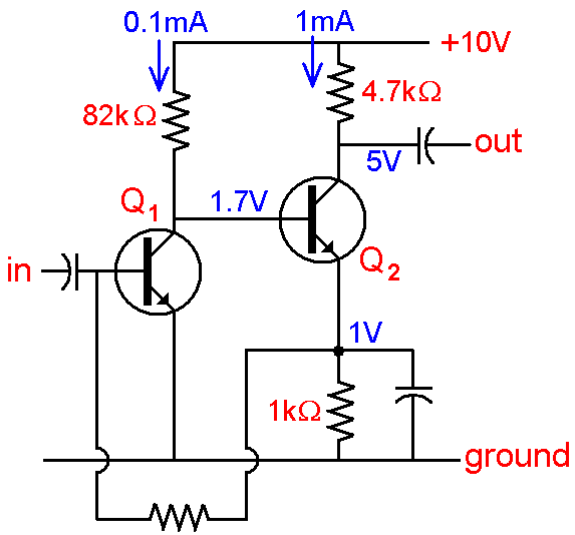


Next, note the capacitor from Q_2's emitter to ground. This eliminates AC degeneration for Q_2 at signal frequencies.

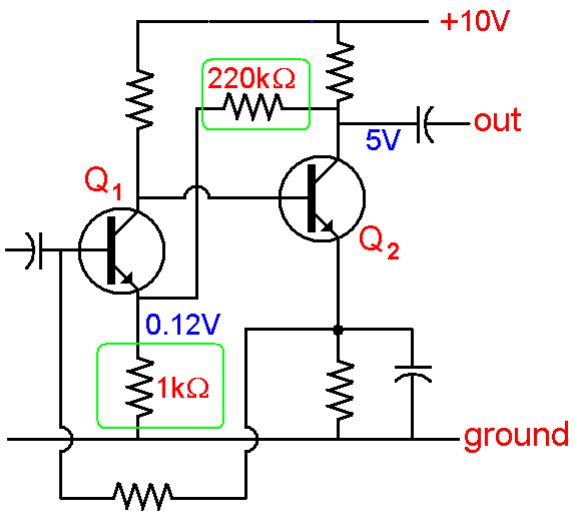
So: V_B of Q_1 is a given, 0.7V. If we make an educated guess that there will be a 0.3V drop across the feedback / bias resistor, that means V_E of Q_2 is 1V, making V_B 1.7V. I'm going to arbitrarily set Q_2's I_C and I_E at 1mA, meaning Q_2's emitter resistor will be 1kΩ. If we have a 10V supply and we follow the normal practice of setting the output transistor's collector at half the supply voltage, that means a 5V drop across the collector resistor. With a 1mA current, that's 5kΩ. 5.1kΩ is the nearest standard value. There's 1V at the emitter, so the collector can't pull all the way down. Maybe we should set the collector voltage a little higher, which we can do by going to a 4.7k resistor.

The choice of I_C for Q_1 is also a bit arbitrary. If Q_2's beta is 100, its base current will be 1mA / 100 = 10µa. Setting Q_1's collector current at 0.1mA insures that there's more than enough current to drive Q_2's base. Often, the choice comes down to noise: for a particular transistor, there's a small range of collector current where least noise is generated. This depends on the transistor and also on the signal source impedance; higher source impedances usually call for lower collector currents.

We've decided Q_2's I_C is 0.1mA and we know its V_C is 1.7V. That makes the drop across the collector resistor 8.3V, which calls for an 83kΩ resistor. 82kΩ is the nearest standard value.



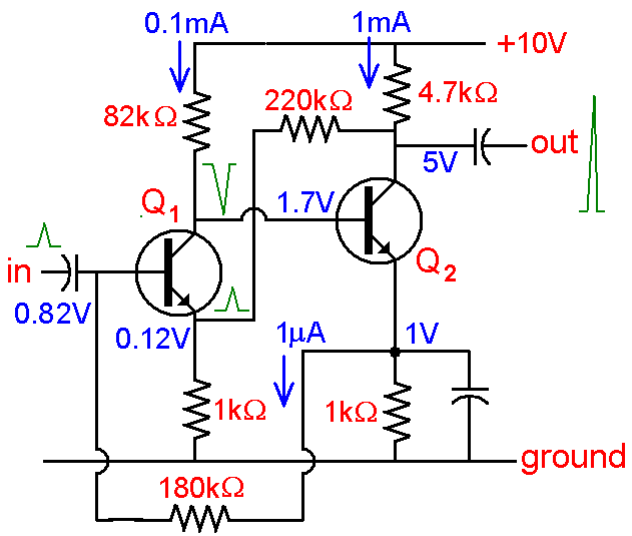
We've just designed an amplifier with lots of gain – too much for most applications. Time to apply some negative feedback. We could put a resistor in series with Q2's bypass capacitor, but we're going to do something different instead. Let's take another look at the complete circuit.



Note the two resistors outlined in green. They form a 221:1 voltage divider that applies feedback from the output to the emitter of Q_1. Recall that, when a signal is applied to the emitter of a transistor stage, it's a common-base amplifier. In this case, Q_1 is a common-base amplifier *from the standpoint of feedback only*.

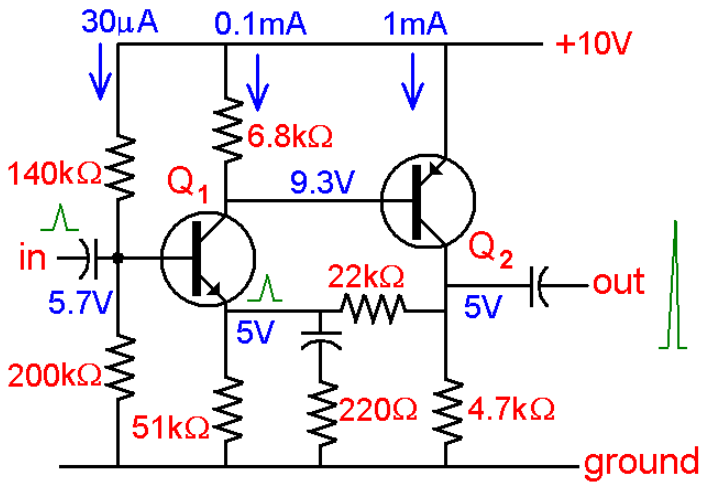
The feedback signal passes through Q_1 without being inverted, then through Q_2 which, being a common-emitter stage, inverts the signal. Thus we have one inversion, and the fed-back signal is negative feedback. Assuming the "natural" gain of the amplifier is a lot higher than the final gain, the feedback loop is the controlling factor. In this amplifier, gain is set for about 200.

Let's figure out the one remaining part, the resistor at the bottom that supplies bias current to the base of Q_1. The 1kΩ resistor in Q_1's emitter circuit changes the operating point a little. There's 0.1mA of current from the transistor and about 0.02mA from the feedback resistor, so the voltage drop across that 1k resistor is about 0.12V. That puts the base at about 0.82V. We assumed the emitter of Q_2 would be at 1V, so that means there's 0.18V across the bottom resistor. If Q_1's beta is 100, base current will be 1μA. 0.18V / 10^{-6}A = 180kΩ. I've added green waveforms to show what would happen to a small positive-going pulse at the input. From a signal standpoint, both stages are CE. That means two inversions, so the output will be non-inverted.



Direct Coupled Example 2

This circuit combines an NPN transistor and a PNP transistor. The two gain stages are CE circuits, so the output is inverted twice, thus it's a non-inverting amp.



Biasing on Q_1 is very strange. Input is to the base and output is from the collector like any other CE amp, but we want the emitter to be at about half the supply voltage, or 5V. That means we need a voltage divider at the input that puts the base at 5.7V. Q_1's output is connected between the base and emitter of Q_2. That means the voltage drop across Q_1's collector resistor has to be 0.7V, which is to say Q_1's collector voltage must be 9.3V.

Q_2 is set up the normal way: its collector is at 5V. Q_1's emitter will also naturally be at 5V, because 0.1mA across a 50kΩ resistor is 5V. (We're using a 51k resistor, the nearest standard value.) So, in the absence of negative feedback, Q_1's emitter and Q_2's collector should naturally be at about the same voltage.

Now we connect the 22kΩ negative feedback resistor. The operating point shouldn't change much, since both ends attached to the resistor were supposed to already be at 5V. Due to normal component variations, they probably weren't actually at the same voltage, but now we have a feedback resistor to bring them closer.

From the Q_1 emitter to the Q_2 collector, there's a signal inversion, so the signal fed back through the 22kΩ resistor is negative feedback. Now we connect a 220Ω shunt resistor to give us an 101:1 voltage divider, setting the overall gain of the circuit at about 100. The capacitor, of course, is there because the 220Ω resistor should not be DC-grounded.

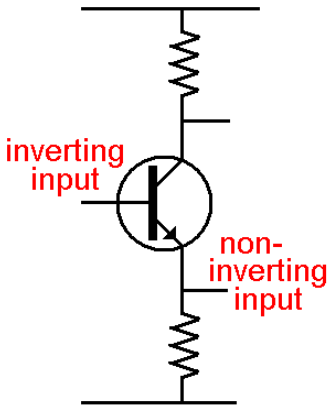
NEGATIVE FEEDBACK'S EFFECT ON OUTPUT IMPEDANCE

As stated back on page 21, negative feedback almost always lowers output impedance. This is because, if a load drags down the AC output voltage, that will also reduce the amount of negative feedback, causing amplifier gain to rise and largely offset the loss. Thus negative feedback not only compensates for nonlinearities and gain variations within the amplifier, it also compensates for varying loads across the output. In an amplifier with several transistors, voltage gain may be very much higher than the final gain after negative feedback is applied. The output impedance of such an amplifier approaches zero ohms.

Negative feedback is like the cruise control of an automobile, and applying a load across the output is like going up a hill. The car will begin to slow down, the cruise control will sense the loss of speed and automatically compensate.

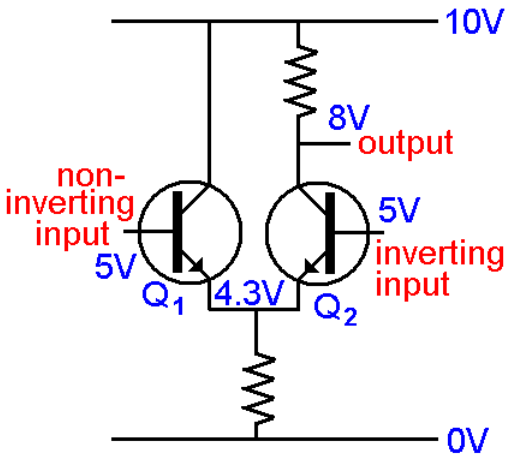
THE DIFFERENTIAL AMPLIFIER

I'd like to call your attention again to the fact that, in the direct-coupled amps above, Q_1 is an inverting CE stage from the standpoint of the input signal, but a non-inverting CB stage from the standpoint of feedback. In other words, it has two "inputs," the base and the emitter, and it is the *voltage difference* across these inputs that creates an output at its collector. If we were to feed exactly the same AC signal voltage to base and emitter, there would be almost no change in collector current and no output.

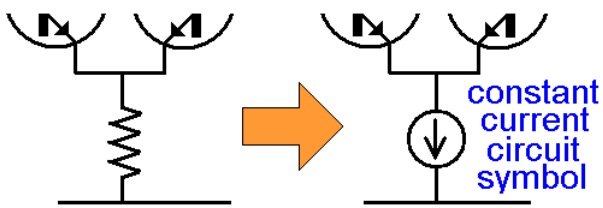


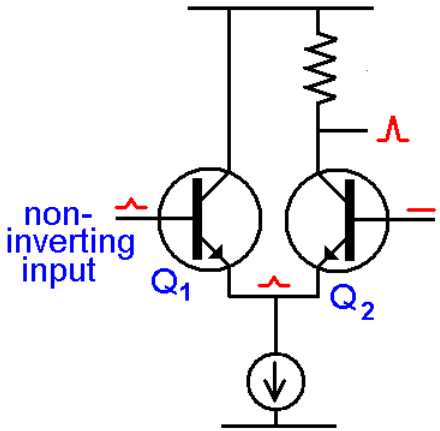
In other words, a transistor can be a **differential amplifier** – one that amplifies a difference across base and emitter, but does not amplify or even pass a signal that is common to both. It's a really crummy "diff" amp, though. For one thing, the impedances at the base and the emitter are nowhere near the same.

A proper differential amplifier is shown above right. Q_1 and Q_2 are joined at the emitters, and if they're both biased on (bias circuit not shown) they will both draw collector current. If the transistors are well matched, the collector currents will be equal. We could say that Q_1 and Q_2 cooperate in holding up the top end of the emitter resistor.



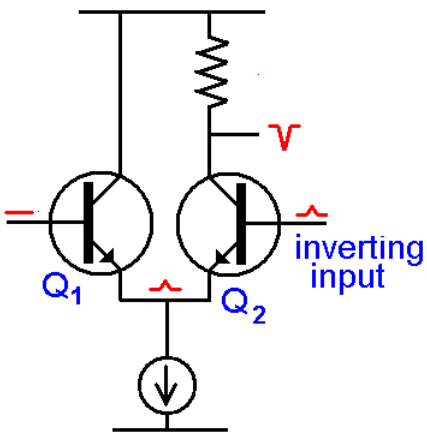
Before we go any farther, let's improve the circuit. Ohm's Law says, if the emitter voltage changes, current through the resistor and thus the emitters will change. It would be nice if total emitter current would stay the same no matter what the emitter voltage is. We can make that happen by replacing the emitter resistor with a **constant-current** circuit, which I will describe shortly. In a constant-current circuit, current flow stays the same no matter how voltage changes.





In the constant current circuit, when the emitters of Q_1 and Q_2 rise or fall, total current does not change. What does happen is that the *balance* of current shifts back and forth like a seesaw.

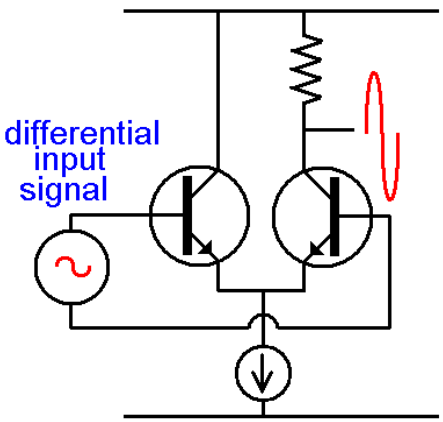
Suppose we apply a small positive pulse, say 10mV, to the base of Q_1. This will cause its emitter/collector current to increase; it will shoulder more of the burden of holding up the emitter resistor, meaning Q_2 will shoulder less. When Q_2 current decreases, voltage drop across its collector resistor decreases, and collector voltage rises. Thus, from input to output, no inversion.



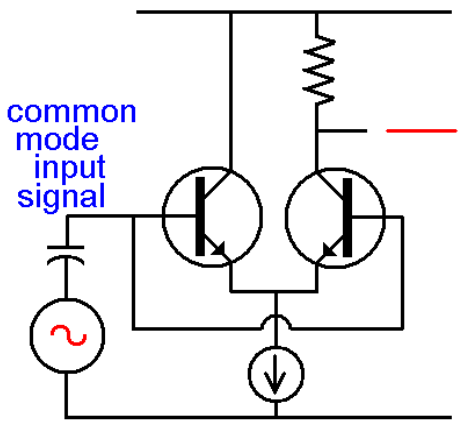
If, on the other hand, we apply a positive pulse to the base of Q_2, it will take part of the combined emitter current away from Q_1. As emitter current rises, collector current increases, while Q_1's collector current falls.

An increase in Q2 collector current will increase the voltage drop across the collector resistor, causing the output voltage to fall, thus from input to output, an inversion.

The differential circuit gives us inverting and non-inverting inputs that are exactly the same, electrically. This enables them to "see" a differential signal but be "blind" to a common-mode signal. Let's apply a differential signal across the two bases . As one V_B goes up and the other goes down, the current balance will shift and an output voltage will be produced at the collector of Q_2.



Finally, let's apply the same signal to both bases (next page). The bases rise and fall together, as do the emitters. Because the emitters' current is set by the constant-current circuit, there will be no change in total current, nor will the current balance between Q_1 and Q_2 shift, so there will be no AC output voltage at the collector of Q_2.



Q_1 and Q_2 are called a **differential pair** or, sometimes, a **long-tailed pair**. This is the starting point for the **operational amplifier**, which is the subject of the next chapter. Since a "diff pair" processes a signal cooperatively, it's considered a single stage.

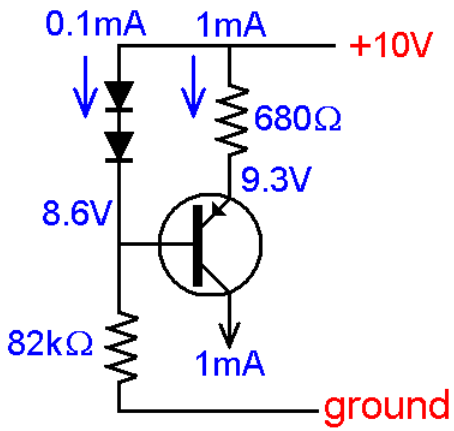
All the multi-transistor circuits I've shown you also work with FETs. The biasing is just different.



CONSTANT-CURRENT CIRCUITS

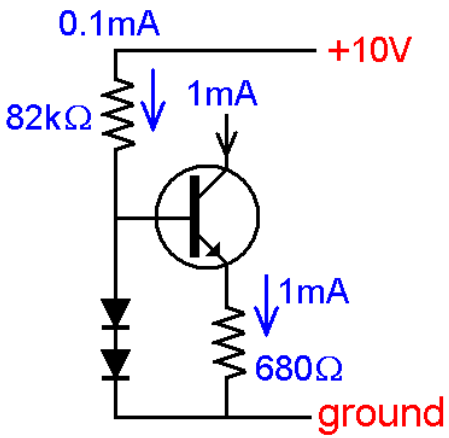
These aren't amplifiers, but I'm going to discuss them here because they're so commonly used in amplifiers. There are many cases where we want to either supply or drain off a constant current, a current that stays the same regardless of voltage – for example, in the diff pair circuit.

A **current source** draws a specified amount of current from a positive rail and passes it to some other circuit. A **current sink** drains a specified amount of current out of some other circuit and flushes it to ground, or where a bipolar supply is used, may flush it to the negative rail. Obviously, whoever assigned these names was thinking in terms of conventional current, flowing plus to minus.



The circuit at right is a **source**. As noted earlier, a transistor's collector current is almost the same as its emitter current. That means, if we use a fixed base voltage and an emitter resistor to set emitter current, we will have set a collector current that doesn't vary with voltage. In this case, voltage between the +DC rail and the base is set to 1.4V by the two diodes. Emitter-to-base voltage is 0.7V; the other 0.7V is dropped across the emitter resistor. We set the emitter current, and thus the collector current, by our choice of emitter resistor. With the values shown, this circuit will supply about 1mA to whatever circuit it's supposed to supply 1mA to.

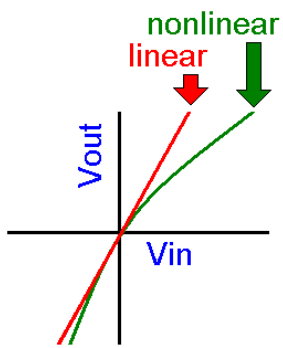
This circuit is a **sink** that drains 1mA to ground. Note that it is exactly the same as the source circuit, except flipped over and with an NPN transistor instead of a PNP.



This circuit would be used in place of the emitter resistor in a differential pair, insuring that a common-mode input voltage will cause no voltage change at the collector of Q_2.

AMPLIFIER LINEARITY

I want to define "linearity" more precisely than I have. If we graph the input signal to an amplifier against the output signal from it – input on the X axis, output on the Y axis – we get a function. Every point along the X axis corresponds to a point along the Y axis, and we can draw a line through all the points to show the function.



Ideally, the function will be a straight line. Then we will have achieved the goal we set out at the beginning, more than forty pages ago: the output signal will be just like the input signal, larger but without any other change. This is what we mean when we say an amplifier is **linear**. No amplifier is perfectly linear, but in the best amps, distortion is 0.001% or less.

Linearity is often very important. In audio, nonlinearity distortion degrades the sound. In an RF amp handling many signals, nonlinearity causes crosstalk between signals and other problems. In a baseband video amp, the transition from black through gray to white, called the grayscale, will not be correct, and nonlinearity also may shift colors, e.g. red may turn into orange.

AMPLIFIER CLASSES

Now that you know how to bias an amplifier so the whole signal waveform can be handled in a linear fashion, I'm going to show you how to screw it up. We're going to talk about amplifiers where current flows during only half the signal waveform, and others where it flows for an even smaller part of the waveform.

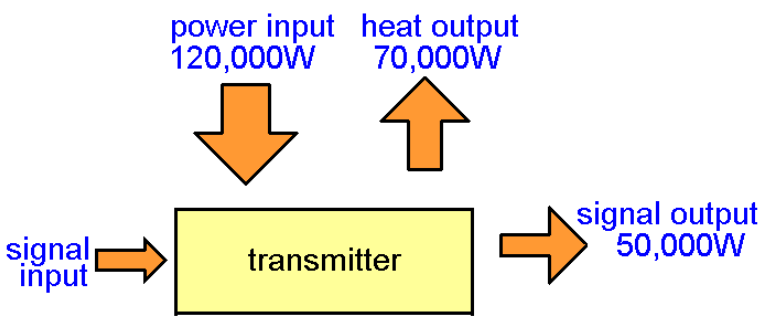
Why on Earth would anybody do that? In a word: efficiency. An amplifier's efficiency is the ratio of the power it puts out as signal versus the power it draws from the power supply, stated as a percentage.

$$\text{Efficiency} = \frac{\text{signal power out}}{\text{supply power in}}$$

The difference is turned into heat.

THE PROBLEM OF WASTED POWER

Fully linear amplifiers, where a transistor conducts all the time, can never be more than 50% efficient, and then only at maximum signal output. Under all other conditions the efficiency is much worse. No big deal at milliwatt power levels, but do we want an AM transmitter that puts 50,000 watts into the antenna and burns up, say, an additional 70,000 as heat? You could power half a dozen houses with that.



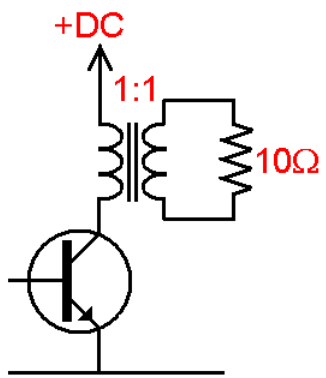
58% of the input power is wasted!

It was the need to improve efficiency that led to the invention of several kinds of amplifiers where the active devices conduct for only part of each cycle.

Amplifiers are grouped into different **classes** according to the percentage of each cycle that current flows through a particular active device.

CLASS A AMPLIFIERS

An amplifier like the ones we have been talking about, where current flows during every part of the signal waveform, is called a **class A** amplifier. I want to spend some time examining a specific amplifier in detail, not just for the sake of understanding the efficiency problem, but also because it will give you a more secure grasp of amplifiers in general.



The circuit at right is the output stage of a transformer-coupled amp that puts out a maximum 10-watt sine wave into a 10Ω load. I've shown a BJT, but the circuit would work the same with an FET.

First we calculate the output voltage we need to put 10 watts into the load. Power into the load, times load resistance, equals voltage across the load, squared.

$$V^2 = P\,R$$

Power is 10 watts, resistance is 10 ohms, so voltage squared is 10 x 10 = 100.

$$V2 = 10 \times 10 = 100$$

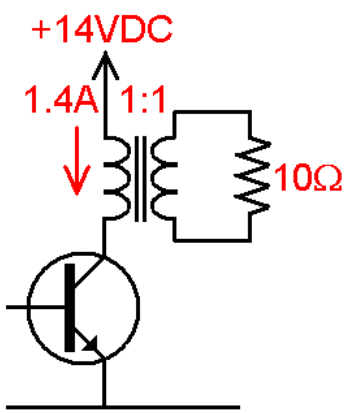
The square root of 100 is 10, so voltage across the load resistor must be 10VAC.

$$100^{0.5} = 10V$$

Next we calculate current through the load

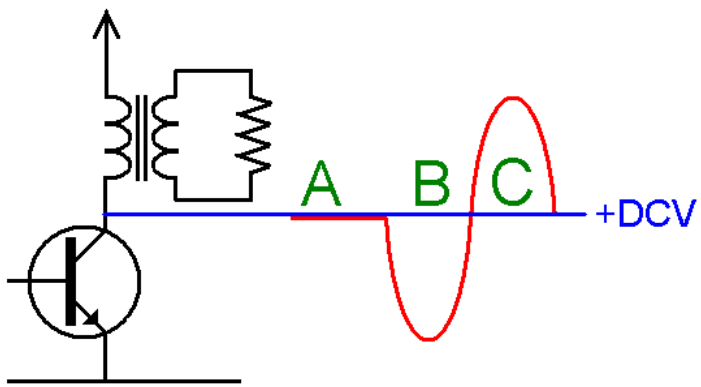
$$10V / 10\Omega = 1A$$

Let's assume we have a 1:1 transformer, so the equivalent load at the collector is also10Ω. The amp will need to produce a 10V RMS sine wave at the collector. That means the peak voltage will be about 14V, or 28V P-P. And the peak current will be 14V / 10Ω = 1.4A.



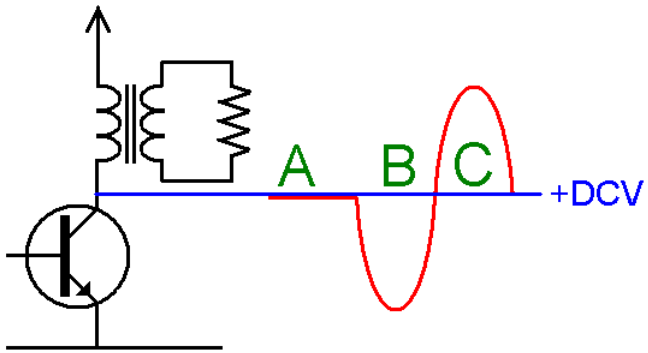
The diagram below left shows the state of the circuit in the absence of signal. To make things simple, we will assume the transformer is ideal, so it has no DC resistance. That means the same 14V will appear at both ends of the primary, and at the collector.

Now let's add signal. In the diagram below, “A” shows the circuit before signal is applied. The collector sits at 14V. 1.4A of current is flowing, and the transformer core is partially magnetized.



On negative alternations, current increases as the transistor pulls the voltage down as at “B” above, and core magnetization rises. At the lowest point in the sinewave, when the transistor is saturated, collector and transformer primary current is 2.8A.

1.4A of this is the original quiescent current. The other 1.4A is current that is transferred through the transformer to the load, as 14V is applied across the 10Ω resistor.



On positive alternations, as shown at “C” above, current falls. Magnetization will fall, creating a reverse voltage across the primary. The falling magnetization will induce a voltage across the secondary, 14V at the peak, which will cause 1.4A to flow through the load resistance. This comes from the stored energy in the transformer. On the primary side, current falls to zero.

Voltage at the transistor end of the primary will go above the power supply voltage during positive alternations. With a loaded transformer secondary, the voltage at the transistor end of the primary may go as high as twice the supply voltage, to +28V. This won't seem strange if you consider that this is the same phenomenon that causes inductive flyback.

Back to efficiency: as we've seen, to accommodate the whole waveform, we need a 14V rail and a quiescent current of 1.4A. In other words, for a maximum output of 10 watts, the circuit must draw 1.4 x 14 = 20 watts. Thus, an ironclad rule of class A amplifiers: with a sinewave signal, efficiency is, at most, 50%.

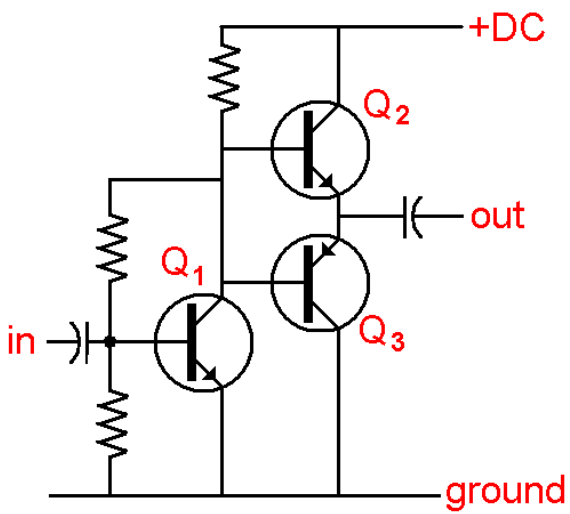
And that's maximum efficiency, when the amp is putting out as much signal as it can. When the signal level is lower – 1 watt, 1/10 watt, zero watts – it's still burning up an unvarying 20 watts. This means not only do we have to supply 20 watts for the sake, often, of 1/10 watt of output, which is bummer enough, but the other 19.9 watts are turned into heat, which we have to get rid of if we don't want the transistor to melt.

CLASS B AMPLIFIERS

A **class B** amplifier is one where current flows for approximately half of each cycle. As such, it isn't very useful. However, if we connect a PNP and an NPN transistor together, so each handles half of every cycle, we get both positive and negative alternations at the output, with a great reduction in wasted power.

The idea is that you can bias a transistor “off,” and let it handle only half of each wave, if there's another transistor to handle the other half.

This simple class B amp is a class A common emitter stage, Q_1, followed by a direct-connected pair of transistors, Q_2 and Q_3. A mirror-image pair of NPN and PNP transistors (or N-channel and P-channel FETS) is called a **complementary pair**.



Q_1 is a common emitter stage that amplifies both current and voltage. It is capable of an output voltage swing that is almost rail-to-rail.

However if it were to put out enough current to drive, say, an 8Ω speaker directly, it would have to draw a lot of quiescent current, and we'd be back to the class-A low efficiency problem. So we allow Q_1 to draw a much lower current, and then connect its output to the bases of Q_2 and Q_3. Q_2 and Q_3 operate in common-collector mode, supplying much more output current than Q_1 alone could, but each one only handles half the waveform. A positive alternation turns Q_2 on and a negative alternation turns Q_3 on. Since the bases are connected together, it is impossible for both base-emitter junctions to be forward biased at once; thus Q_2 and Q_3 cannot both be on at the same time. In the absence of signal, they're both off. The only constant power draw is in the Q_1 class A common emitter stage.

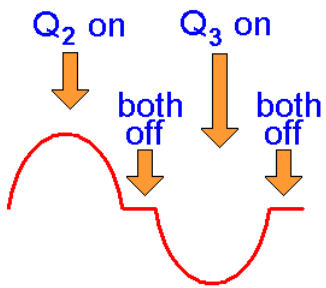
When two transistors alternate like this, each carrying half the cycle, they are said to be operating in **push-pull** mode, though I think “pull-pull” would be more accurate; one transistor pulls the output up toward the DC rail; the other pulls it down toward ground.

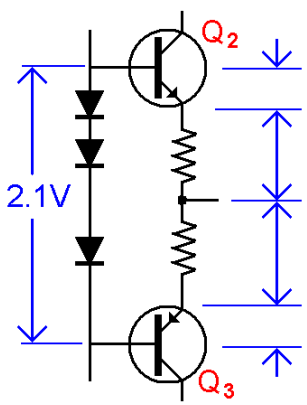
The power savings in class B operation can be enormous. For example, a 100W class A amplifier draws at least 200W at maximum signal output and the same 200W when there's no signal. A 100W amp with a class B output stage draws about 150W at maximum signal output and maybe 5W when there's no signal. That's a 97% power saving under no-signal conditions.

CLASS AB AMPLIFIERS

If we want a really nice undistorted output waveform, the class B amplifier has a problem. There's a glitch in the middle of the wave, due to the fact that there's a 1.4V span where both output transistors are off – 2.8V if they're Darlingtons, 5V to 8V if they're MOSFETs.

Remember, the voltage on the collector of Q_1 has to go up by 0.7V before Q_2 starts conducting and down by 0.7V before Q_3 starts conducting. The distortion caused by the glitches at the waveform zero-crossings is called **crossover distortion.**

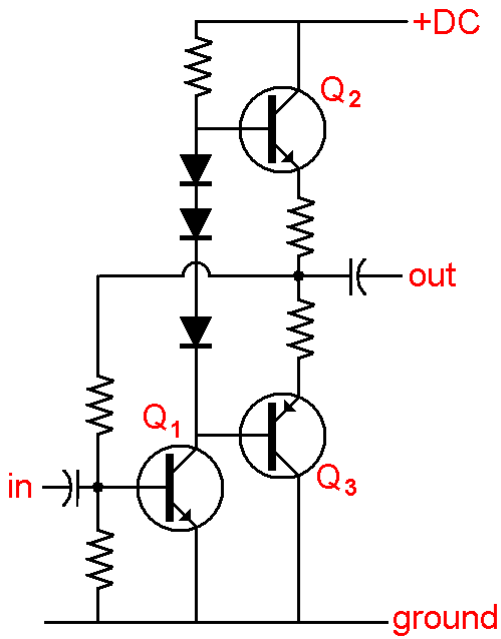




0.7V
0.35V
0.35V
0.7V

We can fix this by biasing Q_2 and Q_3 "on" just a little, so their quiescent collector current is maybe 50mA for an audio power amp. We're making an amplifier that's a cross between class A and class B, commonly known as **class AB**.

For example, if we put three diodes between the bases of Q_2 and Q_3, we'll get an almost unvarying 2.1V drop. The transistor base-emitter junctions drop 0.7V each for a total of 1.4V, leaving an extra 0.7V that appears between their emitters. This voltage is split between the two low-resistance emitter resistors. Thus both transistors are slightly "on."



When a signal comes along, one transistor turns fully on and the other turns off. Half a cycle later, the "on" transistor turns off and the "off" transistor turns on. Note that the negative feedback is now taken from the output, rather than the collector of Q_1. Even with class AB operation, the hand-off from one transistor to the other at zero crossings isn't perfectly smooth, so we need negative feedback to reduce distortion. But what counts is distortion at the output, not at Q_1. So we take the negative feedback from the output.

This circuit, with refinements, would be typical of a medium-quality audio amplifier used in the paging system in a warehouse or department store.

CLASS C AMPLIFIERS

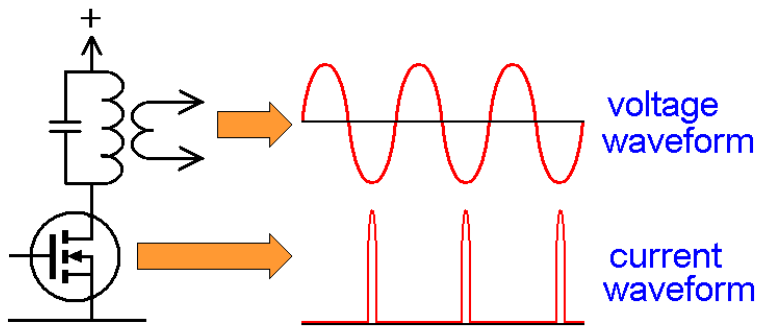
In a class C amplifier, current flows much less than half the time. The class C amplifier is used where waveform distortion is irrelevant. Its primary use is in tuned circuits, like those found in RF transmitters and the RF circuits in radios. The transistor is essentially a switch.

The illustration at right represents a transmitter output. I've shown a MOSFET for the active device because MOSFETs make good switches.

The inductor and capacitor connected to the drain form a resonant tank. Every time the waveform at the drain approaches zero, the transistor turns fully on, shorting the tank to ground for an instant, giving the tank a “kick” to keep the waveform going.

The current waveform in the transistor is nothing like a sine wave, but it doesn't matter: it's the resonant tank that creates the sine wave, and the transistor just injects energy at the right moments.

Since the transistor is a switch, there's very little voltage drop across it when it's on, and no current when it's off, thus it dissipates almost no power. The efficiency of a class C amp can be 90%, meaning an amp putting out 100W may draw as little as 110W from the supply. That's only 10W of wasted power, and only 10W of heat to get rid of.



A, B, AB and C amplifiers are the “traditional” amplifier classes that accounted for almost all amplifiers until about ten years ago.

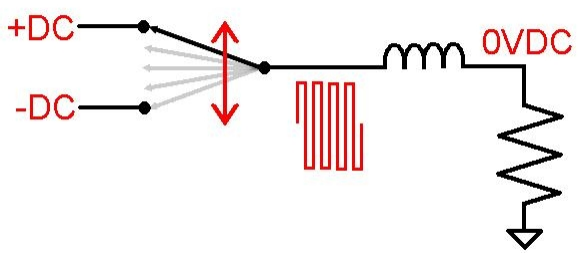
CLASS D AMPLIFIERS

Is there any way we can combine the efficiency of a switched amplifier like class C with the linearity of class A? Today, there is.

This is relatively new technology. The theory has been understood for decades but the components to make it work weren't available until recently.

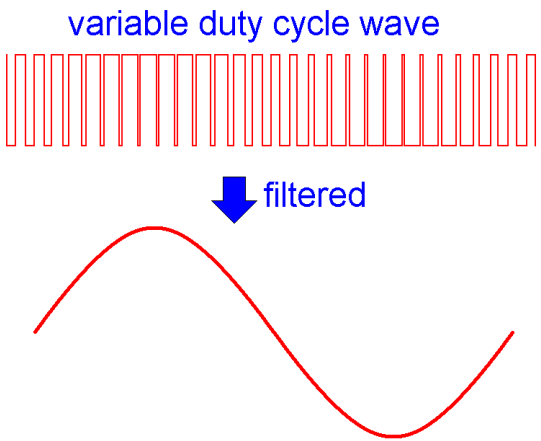
This will be easier to understand if we suppose that, instead of one DC rail, we have two: one that is positive with respect to ground, and another that is negative. Suppose we have a SPDT switch that is constantly toggling between the two at a high frequency. If the switch spends equal amounts of time connected to the plus and minus rails – if the duty cycle is 50% – the average voltage at the output will be zero.

Suppose we connect a load to the switch, with an inductor in the circuit. The inductor will block the high-frequency square wave coming from the switch, and feed only the average to the load.



As long as the duty cycle remains 50%, DC voltage across the load will be zero. AC voltage will be whatever manages to get through the choke, which will be very small if the choke is sized correctly.

If the duty cycle changes, a DC voltage will appear across the load resistor. If, on the other hand, the variation in duty cycle follows a sine pattern, the voltage fed to the load will be a sine wave. The inductor must be sized to block the high frequency rectangular wave but pass the low frequency sinewave.



This is a vastly simplified description of a **class D** amplifier. Getting it to actually work, especially with low distortion, is a challenge. However, it's obvious that this is the future of amplifiers. Class D technology is already dominant in motor control and variable-output power supply applications, and for good reason: class D amplifier efficiency can easily exceed 90%.

Single-Supply Class D Amplifiers

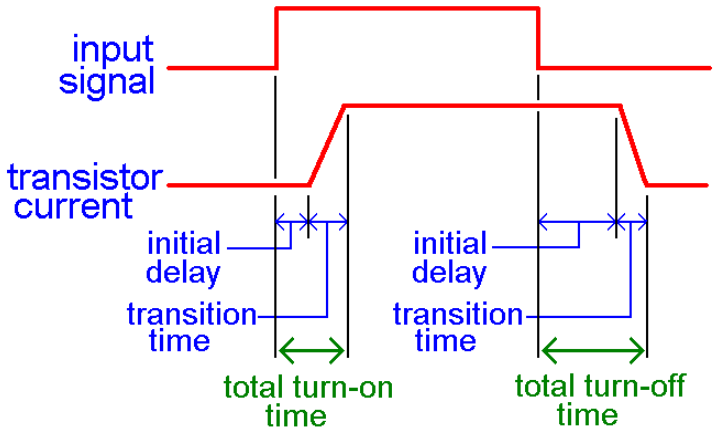
A class D amp does not require a bipolar supply. We can use a single supply rail and ground, in which case the average input and output voltages will not be zero. This problem is solved in the usual manner: we add coupling capacitors to the input and output.

Turn On and Turn Off Times

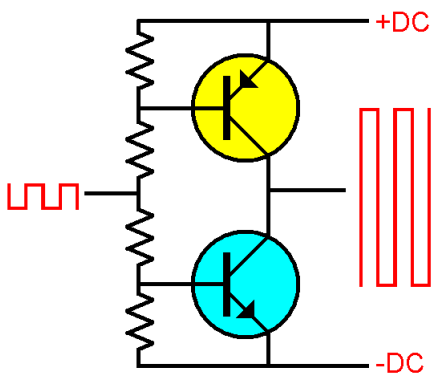
We know that a transistor has a high-frequency limit – everything has – but when a transistor is used as a saturating switch, the problem gets more complex. Switching a transistor on and off involves four time intervals:

- a delay between the time when a turn-on signal is applied to the base and when the transistor actually begins to turn on
- a delay between when the transistor starts to turn on and when it's fully on
- a delay between when the turn-on signal is removed from the base and when it actually begins to turn off
- a delay between when a transistor starts to turn off and when it's fully off

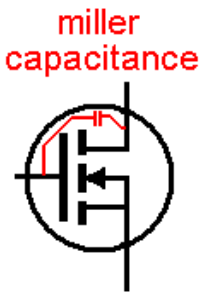
All of these problems are significant, but the worst problem is that it takes longer to turn a transistor off than turn on.



A class D amplifier output is almost always a pair of transistors, one to pull the output positive and the other to pull it negative. If one transistor turns on before the other turns off, there will be a moment when both are on. During this interval, current flows straight through both transistors, essentially shorting out the power supply. This condition is called **shoot-through**.



One of the reasons for this is the internal capacitance between a BJT's collector and base, or a FET's drain and gate, called a **Miller capacitance**. To turn an NPN or N-channel transistor off, the base or gate must go in the negative direction. But, once the transistor starts turning off, the collector/drain goes positive, and the Miller capacitance couples this positive transition back to the input, negative-feedback style, working against turn-off. To get a fast turn-off, it is not enough to stop feeding energy into the input; you must pull energy *out* of the input. You must pull the base or gate down to the emitter or source voltage via a low resistance, or even reverse-bias it.



Class D integrated circuits are now available with complex timing functions that remove the drive to one transistor, wait an instant, then apply drive to the other transistor. If properly adjusted, this takes care of the shoot-through problem. We're going to run into it again, though, in EM4 when we talk about digital electronics.

AMPLIFIER CLASS REVIEW

- Class A: current flows during 100% of the waveform. Good linearity, poor efficiency.
- Class AB: current in each output transistor flows during slightly more 52% of the waveform. Good linearity if carefully designed, better efficiency.
- Class B: current in each output transistor flows during slightly less than 50% of the waveform. Good linearity except at crossover points, better efficiency.
- Class C: current flows during much less than 50% of the waveform. Nonlinear, excellent efficiency.
- Class D: current flows in one or the other output transistor at all times. Good linearity, excellent efficiency, complicated design.

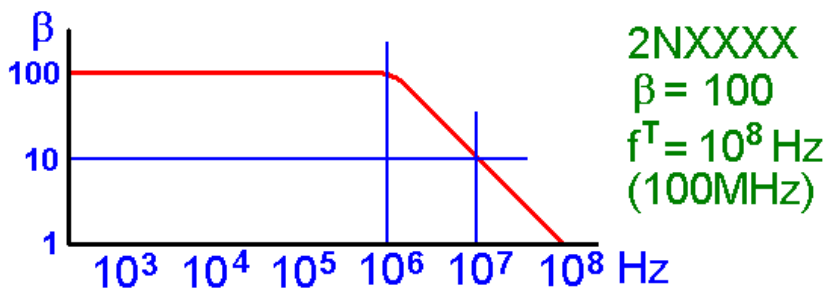
HIGH FREQUENCY LIMITATIONS: f_T, f_{MAX}

Returning to linear, class A operation, there is inevitably a high frequency where an amplifier's gain begins to fall off. It is not possible to specify this frequency for a particular transistor because it depends on too many circuit factors (collector resistor, signal source resistance etc).

Instead, the manufacturer specifies the frequency at which the transistor's beta falls to unity. This is called $\mathbf{f_T}$, the “transition frequency,” which isn't a very useful name. A better and more frequently used name is **gain-bandwidth product**. Let's say a particular transistor's $\mathbf{f_T}$ is 10^8 Hz, meaning at 10^8Hz beta is 1 or unity. f_T is the product of gain times bandwidth, in this case $1 \times 10^8 = 10^8$. Duh.

The beauty of f_T is that it's a constant; it is always the product of gain times frequency. Thus, if f_T is 10^8, then at 5×10^7 Hz, beta will be 2; at 10^7 Hz, beta will be 10, at 10^6 Hz, beta will be 100.

Of course, this only works at high frequencies; it doesn't mean that at 10Hz the beta will be 10^7. As the chart below shows, beta is whatever the manufacturer specifies – in this case, 100 – until we get up to the frequency where the gain-bandwidth product takes over.



f_T is only a ballpark figure. Transistors vary, and circuit details like the collector resistor affect f_T. f_T also varies with current, increasing as current increases up to a point, then falling. Designers choose transistors with f_T much higher than they need so as to stay well away from these problems.

f_T also varies with configuration. It's valid for a common-collector amplifier. Effective f_T is a bit lower for a common-emitter amplifier, because then we have the Miller capacitance to deal with.

For a common base amplifier, current gain is a tiny bit less than unity to start with, so does f_T apply there? Yes, but it's measured differently. f_T is the frequency at which beta falls to 0.707.

It is still possible for a BJT to amplify, even when current gain is 1 or less, because the circuit may have voltage gain. Common-base amplifiers are routinely used at frequencies which are ten times higher than f_T. The frequency where *power* gain drops to unity, beyond which it is impossible to get any gain at all, is $\mathbf{f_{MAX}}$.

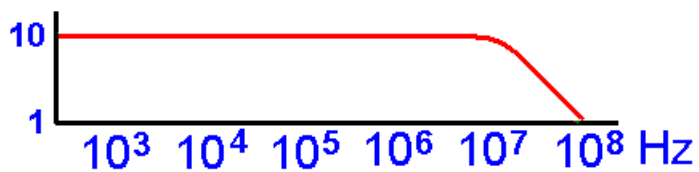
f_T applies differently to FETs, because they're not current-driven. There's a way of specifying f_T having to do with gate-to-source capacitance, but it depends on so many factors besides the FET itself that circuit designers look at other parameters. Sorry, I'm not going there, it would double this already long chapter. Designing high-frequency FET amps is well beyond anything a tech will ever do, but don't let me discourage you from looking this stuff up on the web if you're curious.

PHASE SHIFT AND STABILITY

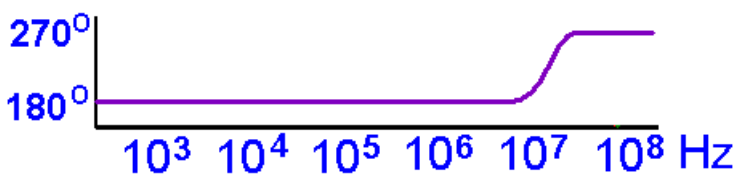
One of my pet peeves is when people say “180 degrees out of phase” when they should say “reverse polarity.” A CE amplifier inverts, meaning it reverses polarity.

However, now that we're talking about high frequencies, I'm going to have to hold my nose and talk about phase. Because, the higher the frequency, the more an amplifier shifts the phase.

The chart below is a CE amplifier whose gain up to 100MHz is controlled by feedback resistors, and above 100MHz is controlled by the gain-bandwidth product.



Let's look at the phase change from input to output. At low frequencies, they're 180 degrees out of phase as you'd expect. But above 100MHz, when response starts to drop off, phase change increases to 270 degrees.



The two things that cause phase shift in a single CE transistor amplifier are the common emitter configuration, which shifts the phase 180°, and the Miller capacitance, which at high frequencies causes a maximum 90° phase shift, for a total of 270 degrees. Nothing to get all het up about.

However, when several stages are involved, phase shift can easily reach 360 degrees, which is the same as zero degrees, at some high frequency. If there's a feedback circuit around those stages, at that high frequency the feedback will be in-phase instead of out-of-phase, and the circuit will go nuts.

Having raised this issue, I'm now going refuse to tell you how to fix it until the next chapter. I am such a jerk. But there's something designers of multistage amps must do to insure stability, and if you run into an unstable amp, maybe that's broken. So that's why I brought this up.

AMPLIFIER APPLICATIONS

There are so many kinds of amplifiers that it's difficult to get a handle on them – but let's try. One way to group them is to list what they amplify:

- Voltage amplifiers boost voltage, but are not expected to deliver a significant amount of power to a load.
- Current amplifiers boost current, but not voltage. They are used where we want to connect an output to a load that is too heavy for it.
- Power amplifiers are designed to drive significant power into a load. Voltage and current amplifiers do boost power, but that's not their primary purpose.

We could also group them according to frequency response:

- Broadband amplifiers are designed to handle a wide range of frequencies.
- Tuned amplifiers are designed to handle just one frequency, or a narrow band of frequencies.
- DC amplifiers, used in control applications, are direct-coupled so they can amplify a DC signal like the output of an electronic thermometer.

Or we could group them according to linearity:

- A linear amplifier produces an output waveform that's a faithful copy of the input waveform.
- A nonlinear amplifier does not necessarily produce a faithful copy of the input waveform, though in a tuned circuit, the output tank will take current pulses from the transistor and reconstruct a sine wave. Sometimes, as in FM radio, the whole point is to drive the amplifier into clipping so as to shear off any amplitude variations.

Any of the following may be built using discrete transistors, integrated circuits, or a combination.

AUDIO AMPLIFIERS

An ideal audio amplifier would perfectly amplify signals over the range of **16Hz to 20kHz**, which is the frequency range

the best human ears can hear. Ideally, there would be absolutely no distortion of the waveform.

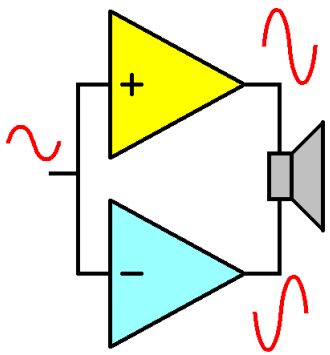
A **line-level** amplifier, used for general signal distribution, typically puts out a maximum of 10V before clipping. Its input impedance is usually 5kΩ or higher, and its output impedance is 2kΩ or lower. An audio amp cannot usually be loaded with its own impedance – it cannot be used in a matched system. Consumer equipment line outputs should not be loaded below 5kΩ and pro equipment line outputs should not be loaded below 600Ω.

The best line-level amps are nearly perfect, and even inexpensive ones are pretty good. With today's technology, it isn't much of a challenge.

An audio **power amplifier** is designed to be connected to a standard 8Ω speaker, or in some cases, two or four 8Ω speakers in parallel, for a combined load of 4Ω or 2Ω. The actual output impedance of the amp is usually a very small fraction of an ohm, but again, you cannot connect it to a load of the same impedance.

It is often desirable to sum both outputs of a stereo amp into one speaker, but they cannot be paralleled. If the two gains differ by even a tiny amount – and they always do – the outputs will fight each other.

Instead, we use a **bridging** connection. Both inputs are fed the same signal, but the polarity to one of them is reversed. This means the outputs will be of opposite polarities; when the left channel output goes positive, the right channel will go negative. The speaker is connected across the "hot" terminals of the two outputs. In theory, this doubles the voltage into the speaker, compared to a single output, and quadruples the power. In reality, when you draw more power, the power supply sags a bit, so you don't quite get what theory would predict.



It is always necessary to check the manufacturer's spec sheet to be sure the amplifier can handle this kind of use. The proper speaker load will be different for a bridged amplifier than for the same amplifier connected the usual way. Many power amps are built with an internal circuit that flips the polarity to one input, so the user doesn't have to worry about it.

There is another, entirely unrelated usage of the word "bridging." When you connect a light, high-impedance load to an output, so the output is not significantly loaded, this is called a **bridging** load as opposed to a **matching** or **terminating** load.
Nowadays, in line level audio circuits, this is the normal kind of load. Microphones are also designed to be bridged, not matched. When someone says "bridging," it's usually obvious from context which kind of bridging they mean.

It's difficult to design a power amp that's as good as the average preamp, due to the complication of a class AB output and the fact that a speaker's impedance varies with frequency. You can get a pretty good stereo receiver for $150, but if you want better, each tiny improvement is exponentially more expensive. A high-end power amp costs $10,000 or more – often a lot more.

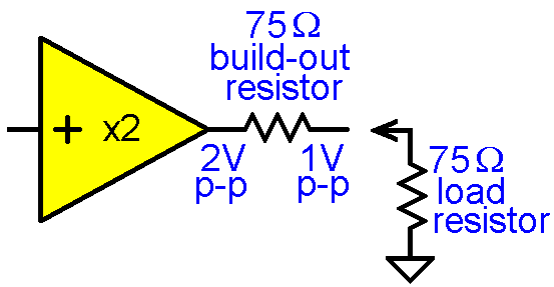
A **preamplifier** is used to boost a very weak signal up to line level. For example, a microphone's output is often on the order of 1mV, meaning that it must be boosted around 1,000 times or 30dB to reach line level. Since many low-voltage signal sources vary a great deal in strength, it is often necessary for a preamp to have a gain control, even if there is another gain control elsewhere in the chain.

The primary measure of a preamp's quality is not low distortion, which is no challenge these days, but low noise. Ideally, preamp noise should be so low that the noise already present in the signal is the dominant noise source.

VIDEO AMPLIFIERS

The expression “video amplifier” usually refers to an amp designed for **baseband** video, video as it comes out of the camera. The main requirements for a video amplifier are a very wide, flat response, 75Ω input and output to match RG59 cables, and noninversion. Standard-definition video requires a bandwidth from 30Hz to at least 4MHz; hi-def, 100MHz or more depending on format. There is almost no other place in analog technology where such a wide bandwidth is required.

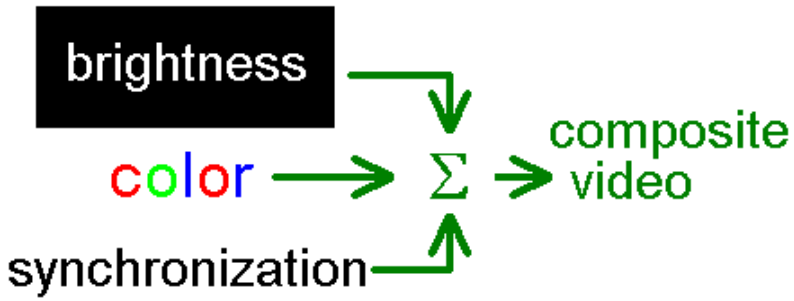
Peak-to-peak signal is around 1 volt. Usually, the 75Ω input impedance is achieved by shunting a higher-impedance input with a 75Ω resistor, and the 75Ω output impedance is achieved by combining an amplifier that has a very low output impedance with a 75Ω series resistor. This means the amp must actually put out 2V, half of which is dropped across the series resistor and half across the load.



In a video amplifier, polarity must be maintained, as instantaneous voltage corresponds to brightness of a point on the display. An inverted signal would look like a photographic negative. Also, the picture would probably be unstable, because the display device is looking for negative-going synchronization pulses and an inverted signal would have positive-going sync.

Thus an inverting amplifier cannot be used, meaning that a video amp cannot be a single-stage CE / CS amplifier. It cannot be a CC / CD amplifier either, because voltage gain is necessary – remember that 2V output – and in a CB stage the gains and impedances don't work out right. A video amp, then, must have at least two stages. Today, there are so many good integrated circuit video amps that few people build video amps using discrete transistors.

In traditional video, separate brightness (**luminance**), color (**chrominance**) and synchronization signals are combined into one signal called **composite video**. (Σ is the Greek letter Sigma, the mathematical symbol for "the sum of.")



That was quite a spectacular feat of engineering back in the 1950s, especially since the color system had to be compatible with the existing black & white standard. The result was a kind of shotgun marriage of a lot of different signals, with some very odd results. For example, if a TV personality wears a black-and-white shirt with a busy pattern, it will show up on most TV's with weird, shifting colors reminiscent of a 1960's drug trip. Sportscasters seem to do this deliberately. And some colors are problematic. Purple can look blue on one monitor and brick-red on another.

It's possible to get a better picture by un-bundling some of these signals and routing them over separate wires. A simple option is to send the color information and the luminance information separately. This is called the **S-video** format. **VGA** and its higher-definition siblings like **XGA** send five separate signals: red, green, blue, horizontal sync, vertical sync. There is no luminance signal, as luminance is simply the sum of the three primary colors.

When signals are separated this way, a new issue arises: timing differences. If the time delays throughout the system are not precisely the same, the red, green and blue signals will not combine properly on the display, and images will have colored edges. Even a few inches' difference between the cable lengths can cause noticeable color fringes.

The amplifiers' frequency response and phase behaviors must be closely matched. The usual way of doing this is to simply design the amplifiers with much wider response than required, e.g. if our red, green and blue signal spectrum extends to 10MHz, use amplifiers whose frequency response is linear to 50MHz.

RADIO FREQUENCY AMPLIFIERS

RF amplifiers can be divided into untuned, broadband amps and tuned, narrow-band amps.

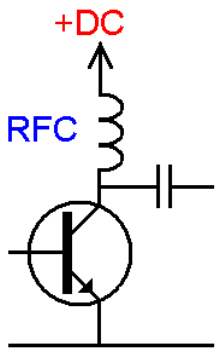
Broadband RF Amplifiers

Broadband amps are linear amplifiers that amplify a wide range of frequencies. They're a lot like video amps except that they don't have to handle low frequencies. The lowest frequencies they handle are still RF. Among other things, this means if we want to use transformer coupling, we can. RF transformers are small and cheap.

In an amplifier that handles many separate signals at different frequencies, linearity is very important. A nonlinear amplifier will **intermodulate** the signals, causing crosstalk between channels and various odd and hard to pin down problems.

An RF amplifier stage inside equipment may have any input and output impedance, but where it connects to a transmission line, its impedance will usually be 50Ω. The exception is video systems, which are 75Ω. When we talk about video in this context, we're talking about **modulated** video, which is a baseband signal that's been modulated onto an RF **carrier**. For example, for the traditional TV channel 2, the 30Hz-4MHz baseband video signal, plus audio, is turned into a 6MHz-wide RF signal that occupies the band between 54MHz and 60MHz. Thus the amplifier that would handle a modulated TV signal is simply a standard wideband RF amplifier, but with input and output impedance of 75Ω instead of 50Ω.

I've mentioned that a transistor amp's gain can be improved by replacing the collector resistor with a choke. This is very frequently done in RF amps, where the frequencies are high enough that the choke can be small and cheap.



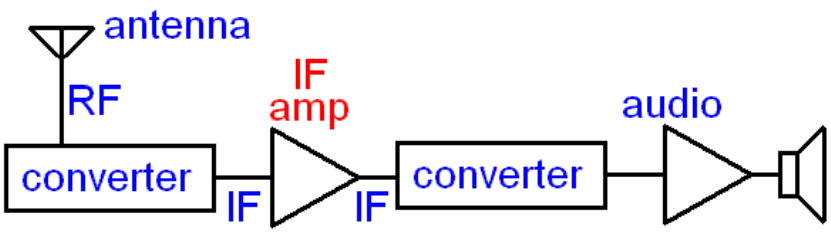
Tuned RF Amplifiers

A tuned RF amp is designed to amplify a single frequency, or a narrow band centered on a single frequency. It has a resonant tank in its output circuit and perhaps tuned components elsewhere. If the signal is a single-frequency sine wave, the amplifier can operate class C. The output tank will take this pulsed energy and reconstruct the sine wave.

INTERMEDIATE FREQUENCY AMPLIFIERS

A radio can be designed to convert an RF signal directly to audio, baseband video, or whatever kind of signal the RF is carrying, and the earliest radios were designed that way.

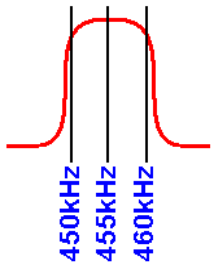
Since the 1930s, a more common arrangement has been to convert the incoming RF to an intermediate-frequency or IF signal. For example, when you tune in a station on an AM radio, no matter what frequency you tune to, it's converted to 455kHz.



This is an “intermediate” signal between what comes in at the antenna and what comes out of the speaker. It's still RF, so an IF amp uses RF design techniques.

In EM6, we're going to study IF amps in detail, but here's the nickel version. Selecting one signal and rejecting all others requires extreme filter circuitry. If we had to retune all that filter stuff every time we changed the station, the process would become impossibly complex. However, if we can convert any incoming radio frequency to a predetermined intermediate frequency, we can make the filters as complicated as we need to, and only tune them once, at the factory.

The most important characteristic of an IF amp, then, is complex filtering to achieve a frequency response that passes frequencies within a well defined range and rejects frequencies outside that range. The range depends on the application; for example, with AM radio, it's 450-460kHz; for FM it's 10.6 to 10.8Mhz.



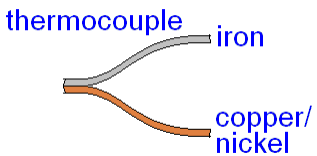
OVERDRIVEN AMPLIFIERS

In some applications, the only thing that matters about an RF signal is its frequency. Amplitude variations are not only irrelevant, they amount to noise. These variations can be eliminated by driving an amplifier into clipping, or in the case of a tuned class C amplifier, by driving a transistor hard enough to saturate on every cycle. This is often called a **limiting amplifier**. This is an important part of the operating principle behind **frequency modulation** or **FM** radio, which we'll examine in detail in EM6.

DC AMPLIFIERS

Up to this point, I've been talking as if “signal” always meant AC. That's not necessarily the case. A signal is any electrical transmission that carries information.

The oldest electrical signal was probably the doorbell, which originally operated on a DC battery. A telegraph is DC. And there are many cases where we must amplify an analog DC signal.



For example, a **thermocouple** is a junction of two dissimilar metals which puts out a voltage that's proportional to temperature. A large stack of thermocouples is sometimes used to put out significant power, as NASA does in spacecraft. But, when a thermocouple is used as a temperature sensor, it's more practical to use a single thermocouple and then amplify its output voltage with a DC amplifier.

In one way, a DC amplifier is simpler than many AC amplifiers because it doesn't have to handle high frequencies. In another way, it's much more complicated. The precise DC operating point of each stage, which doesn't matter with an AC amp, is critical with a DC amp. And all the stages must be direct-coupled, meaning that any error in the DC operating point of the first stage will be amplified.

Design of discrete-transistor DC amps is quite a science, but it's an obsolete science. Nowadays, we make DC amps using integrated-circuit **operational amplifiers**. What's an operational amplifier? It's the next chapter, that's what. So, for the moment, I'm going to put DC amplifiers on the back burner and return to them at the end of chapter 4.

After 62 pages of transistor amps, I'd bet a large sum (if I had a large sum, which I don't) that your brain is mush. So go ahead to chapter 4, but keep coming back here until you have every concept nailed down.

End of chapter 3

EM 1 Chapter 4
OPERATIONAL AMPLIFIERS AND DC AMPLIFIERS

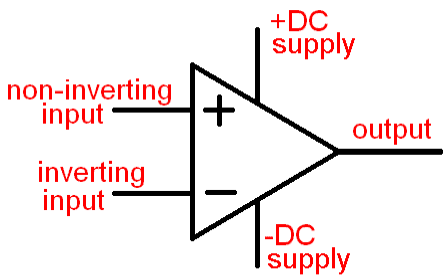
Objectives of this chapter:

- *define "operational amplifier" and describe its input and output functions*
- *give the gain formula for the opamp itself*
- *describe how gain is set in an opamp circuit, and the function of negative feedback*
- *explain why opamps are typically made with extremely high gain*
- *give the gain formula for the noninverting amplifier configuration*
- *give the gain formula for the inverting amplifier configuration*
- *explain why gain must fall off at high frequencies*
- *give the condition for stability at the high frequency where negative feedback becomes positive feedback*
- *describe input and output impedances in inverting and noninverting amplifiers*
- *describe the function of a summing amplifier*
- *explain why the transistors in an opamp are always well matched*
- *describe required precautions to avoid opamp damage*

THE OPERATIONAL AMPLIFIER

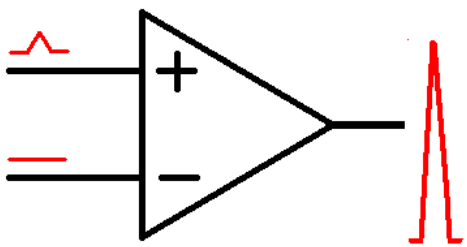
An **operational amplifier**, whose abbreviation is variously spelled **op amp**, **op-amp** or **opamp**, is a special kind of high-gain amplifier with a differential input and a single output. The first opamp, in the late 1930s, was a vacuum tube unit. Around 1960, discrete-transistor opamps began to be made. Today, almost all opamps are integrated circuits.

An opamp has a **differential** input. To review from chapter 3, a differential input has two terminals. The opamp regards the voltage difference *between* these terminals as the input voltage. The absolute input voltages are irrelevant. If both inputs are exactly 0V, +5V, or -5V, the output of an ideal opamp would be 0V in each case. The opamp only responds to a voltage difference.

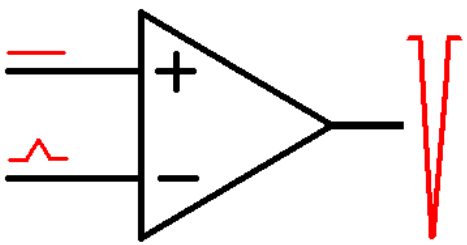


(DC connections are usually not shown in schematic diagrams)

We can look at each of the inputs separately, by holding the voltage at the other input constant. The input labeled "**+**" is the **non-inverting** input. If we change the voltage on this input relative to the other input, the output voltage will change in the same direction; i.e. if the input voltage goes up, the output voltage goes up. The opamp does not invert the input signal. From a differential standpoint, we created a positive difference across the inputs, so the output went in the positive direction.



The other input, labeled "**-**" is the **inverting** input. If we change the voltage on this input relative to the other input, the output voltage will change in the opposite direction; i.e. if the input voltage goes up, the output voltage goes down. The "-" sign can be thought of a reversal, just as a banker regards negative profit as loss. From a differential standpoint, we created a negative difference across the inputs, so the output went in the negative direction.



OPAMP GAIN

Most of the time, when an opamp is used, the reason is not to obtain a differential input. Opamps usually deal with single-ended (non-differential) signals. The real reason is that an opamp offers an amp in a small, convenient package whose gain is completely under the user's control.

With opamps, the kind of gain we're looking for is almost always voltage gain. The gain equation for an opamp is:

$$V_O = (V_N - V_I)G_V$$

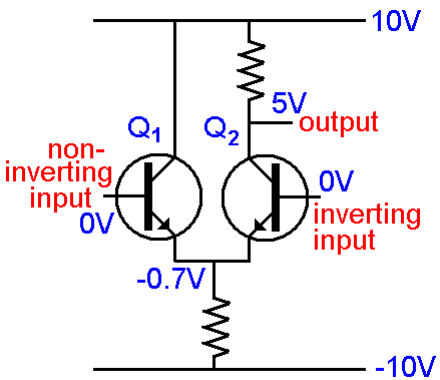
where V_O is voltage output, V_N is the voltage at its noninverting input, V_I is the voltage at its inverting input, and G_V is the "raw" voltage gain of the opamp. In plain English, the output voltage is the input voltage difference multiplied by the opamp's voltage gain.

It is important to understand that the behavior of an opamp by itself is different from its behavior in a circuit. The opamp is a general-purpose **gain block** whose behavior in a circuit is *defined by the parts around it*. It can can amplify with inversion, amplify without inversion, amplify some frequencies and not others, rectify an AC signal, modify waveshapes, detect thresholds, oscillate, create pulses of specified duration, and on and on.

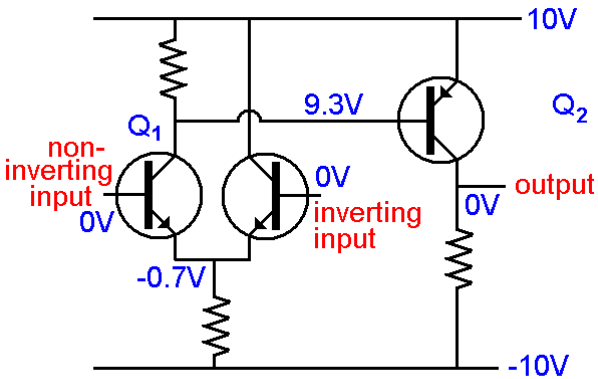
Mathematicians, who like to view things in broad terms, describe these functions in terms of mathematical operations, thus “operational amplifier.” The thing that makes opamps so useful is that their behavior in a circuit is not defined by the manufacturer, but by the user.

The following example circuits have bipolar power supplies, with both positive and negative supply rails. This allows the inputs and outputs to rest at zero volts, and swing both positive and negative from there. I'm doing this just for convenient analysis; an opamp circuit does not need to be powered by a bipolar supply, though many are. To avoid complicated diagrams, I will usually not show the power supplies. Just assume that each of the amplifier triangles has two invisible connections. One goes to a +10V rail; the other to a -10V rail.

A SIMPLE DISCRETE-TRANSISTOR OPAMP



If we try to use the differential amplifier we studied in chapter 3 as an opamp, we run into a problem: the DC collector voltages must be much higher than base voltages, meaning the DC voltage at the output has to be considerably higher than the DC voltages at the inputs. In the circuit below left, if we want the inputs to be at 0VDC (and we do, you'll see why in a bit) the output must be quite a bit higher to keep Q_2 from saturating. If we want to apply negative feedback between the output and inverting input, we have a DC voltage difference to deal with. This complicates negative feedback and all but eliminates the use of differential amplifiers as DC amps. Also, the maximum output voltage swing is limited.

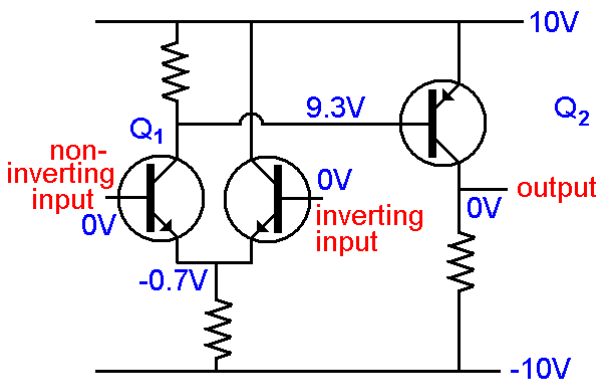


We can fix this problem by adding a PNP transistor. Now the output is the collector of Q_3, not Q_2. It can be at 0VDC, making both inputs and the output the same DC voltage. Also, the output can now swing almost from rail to rail.

Note that the collector resistor for the diff pair has been moved over to the collector of Q_1. Q_3 is a CE stage, which inverts the signal. If we want to keep the non-inverting input on the left, Q_1 has to invert the signal so that, when Q_3 inverts it again, it comes out non-inverted.

And why would we want to keep the non-inverting input on the left? Because the inverting input is where negative feedback is applied, which comes from Q_3's collector. Just a way of drawing the circuit so it lays out neatly on the page.

Typically, DC bias voltage is midway between the supply rails, and is applied to Q_1. If we use a bipolar supply, the bias voltage at the base of Q_1 can be zero.



Here's the diagram again, take your finger and trace a signal through the circuit. If we apply a positive-going pulse to the non-inverting input, Q_1's collector current will increase. This will pull on the base of Q_3, increasing its collector current and pulling up on its collector resistor – so the output voltage will go up.

If we apply a positive-going pulse to the inverting input, Q_2 will draw more current, meaning Q_1 will draw less, which will decrease Q_3's base current, which will decrease Q_3's collector current. The output voltage will go down.

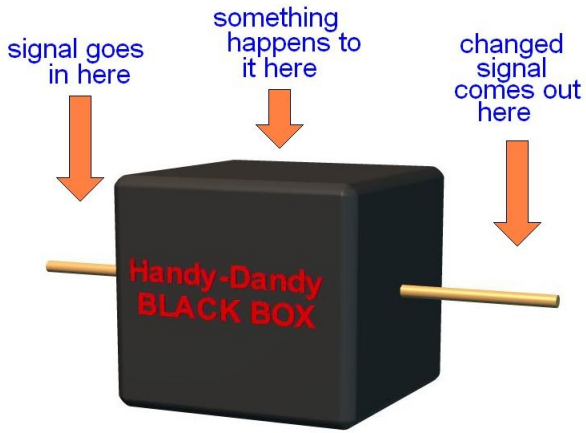
There's one place where this circuit is used all the time: in audio power amps, with the addition of a complementary output pair of high power output transistors. But in our discussion, we'll be using integrated opamps from here on.

THE BLACK BOX CONCEPT

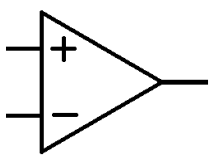
Most opamps nowadays are built as **integrated circuits** or **IC's**, which we'll study in more detail in the next chapter, where many transistors are contained in a single semiconductor package. There are hundreds of integrated opamps available. Many can do a better job than a discrete opamp would, and they're all a lot simpler to use.

The internal circuits of most analog IC's are a lot more complicated than most discrete circuits are. After all, an IC is essentially a photograph. The whole circuit is created together, a layer at a time, by a photographic process, like making an 8x10 glossy of Great-Aunt Fenestrude. Just as a photo of Auntie Fenny alone or a photo of her whole family costs the same, an IC costs the same whether it has a few or a lot of components, as long as the extra stuff doesn't make the chip significantly bigger or require extra processing steps. This is one of the IC's great virtues: if more transistors can give us better performance, we can have them at little or no cost.

How does a tech deal with this extra complexity? A tech doesn't. An IC is designed to be treated as a **black box** whose internal circuit is a mystery. You only need to know what it does, not how. Good engineers, and the very best techs, actually do study diagrams of what's inside the IC's they use. It often helps maximize performance or head off esoteric problems. For a student, it's a good exercise; by tracing biasing and signal flow you can learn a lot. But it's also okay to just treat ICs as black boxes. Manufacturers design them to be used that way.



I showed you a discrete-transistor opamp on the previous page so you could see the concept, but from here on out we will be dealing with the opamp as a black box; that is, as a triangle with three terminals: an inverting input, a non-inverting input, and an output. When an opamp is used in a circuit, that's the way the schematic diagram shows it, usually with the opamp's DC connections omitted to avoid clutter. (The DC connections are called out somewhere else in the diagram.)



OPAMPS AND NEGATIVE FEEDBACK

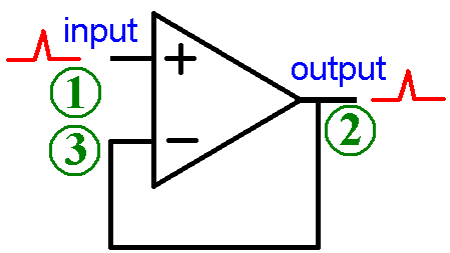
In the inverting and non-inverting diagrams I showed you on page 2, the signal waveforms would suggest an opamp with a voltage gain of about 10. In an actual opamp, the "native" voltage gain is very high, typically 200,000x, which is way too much gain to be useful. Suppose we set up an opamp circuit with both inputs and the output at zero volts. Suppose we then apply a tiny +1mV pulse to the "+" input. Since this input is non-inverting, the output will go positive like the input. If the voltage gain is 200,000 times, the output will try to go to +200V. What will actually happen, of course, is that the output will go positive until it hits the positive supply rail. Even a 100μv input pulse is likely to make the output hit the supply rail. It would seem that an amp with a gain of 200,000 is ridiculous, but it isn't if you understand what's going on.

In chapter 3, we saw how negative feedback can be used to stabilize a circuit and reduce distortion, at the cost of lowering gain. Opamps carry this concept to its logical conclusion: the "naked" opamp gain is so high that it ceases to have much effect on final circuit performance. Actual circuit gain is almost completely determined by the feedback network we connect to it. And, since those parts are chosen by the user, not the opamp's manufacturer, the final gain can (within limits) be anything the user wants.

Here's how it works: as stated above, the opamp's output voltage is the voltage difference across the inputs, times its voltage gain. When we apply a signal to an input, the opamp, via the negative feedback network, drives the differential voltage across the inputs to near zero. It is the design of the feedback network that determines what the output must do to make this happen, which is another way of saying that operation is controlled by the feedback network.

THE VOLTAGE FOLLOWER

Suppose, for example, that the output is connected directly to the "-" input, so we have 100% negative feedback. Obviously, in this circuit, the output voltage and the voltage at the "-" input will always be the same, since they're connected together.



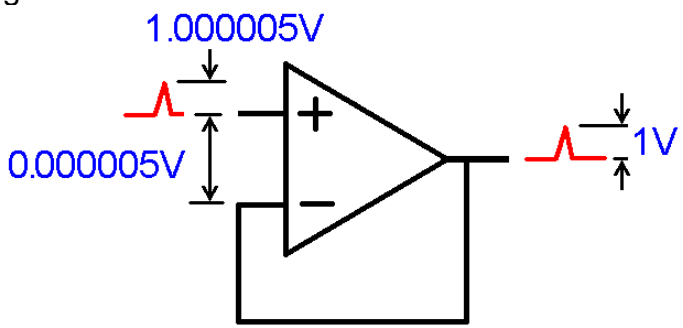
Suppose we apply a +1V pulse to the "+" input. Initially, this will cause a 1V difference across the inputs, and the output will start to go positive. Without the feedback circuit, the output would rise until it hit the positive rail. But with the feedback circuit, as the output rises, the voltage at the "-" also rises. So, we have a sequence: (1) the input rises, (2) the output rises (3) the differential across the inputs is driven to about zero, at which point the output sees no differential and so stops rising. In a fairly slow opamp, this happens in a microsecond.

Input voltage: 1V. Output voltage: 1V. Thus our opamp's monster gain has been reduced to unity. We've just made an opamp voltage follower that behaves about like a common-collector voltage follower, but is better in many ways. For one thing, most opamps have a push-pull output circuit, meaning they can pull the output positive and negative with about equal strength (unlike the CC follower, where the component pulling one direction is a resistor). For another, it doesn't have that 0.7V offset.

Though it's customary to pretend that the feedback drives the voltage across the inputs to zero, of course it doesn't really; if there were not a small voltage across those inputs, the output would not be positive or negative. But we're talking a reeeeeeely small difference voltage! I'm going to flog this horse for a while because it's important for you to fully understand what's going on.

In the example, the output is 1V. If the "naked" opamp's gain is 200,000, a 1V output will be created by an input differential of 1V / 200,000 = 5 x 10^{-6}V. That is, if the "+" input is more positive than the "-" input by five microvolts, the output will be positive by 1V.

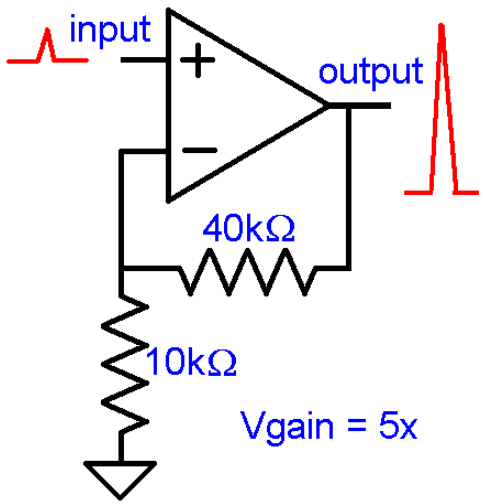
To make the math a little easier, I'm going to start by applying a +1.000005V pulse to the "+" input. The output will start to rise. When it gets to 1V, it will stop rising. That's the point where the voltage across the inputs is 0.000005V. If it went any higher, that would reduce the voltage across the inputs, which would start lowering the output voltage, bringing it back to 1V.



You can see why the actual differential across the inputs is usually ignored. Only in the most exacting measurement situations is the error significant, or even detectable.

THE NONINVERTING AMPLIFIER

Now let's replace the direct feedback connection with a voltage divider consisting of a 40kΩ resistor and a 10kΩ resistor, which is a 5:1 divider. We start with a +1V pulse at the "+" input. The output starts to rise, but will rise only to the point where the "-" input is 1V. With the follower circuit, that would imply 1V at the output. However, in this case there's a 5:1 voltage divider between the output and the "-" input, so the output has to rise 5V for every 1V of rise at the "-" input. Thus, we have an amplifier where 1V of input gets you 5V of output, or an amp with a voltage gain of 5.



Here we have an opamp with a native gain of 200,000, but the final circuit has a gain of 5. Obviously, the controlling factor is not the opamp gain, but the feedback network.

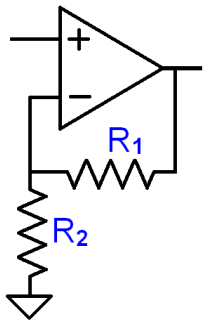
In addition to flexibility and predictability, a benefit of a feedback-controlled amplifier is low distortion. A resistor is almost perfectly linear, at least at low frequencies. Assuming the correct opamp is chosen, so the resistors really do control circuit operation, signal distortion can be reduced to thousandths of one percent.

Just to make sure this feedback-control concept is locked in, let's go over it one more time.

- In normal operation, the voltage difference across the inputs is always approximately zero.
- When a signal is applied to the “+” input, the opamp responds by changing the voltage at the output so as to drive the “-” to the same voltage as the “+” input.
- It is the output voltage that causes the “-” input voltage to change, via the negative feedback path.
- How far the output has to change to drive the input differential to zero depends on the voltage divider ratio of the negative feedback circuit.
- Thus the circuit gain is determined by the voltage ratio of the feedback divider, not the gain of the opamp.

At the risk of getting ahead of ourselves, I'll mention that if the feedback network includes capacitors and/or inductors, the voltage divider ratio will vary with frequency. This permits us to create an amplifier with different gain at different frequencies. Or, we can put a diode in the feedback path and create an amplifier that rectifies. The possibilities are almost endless.

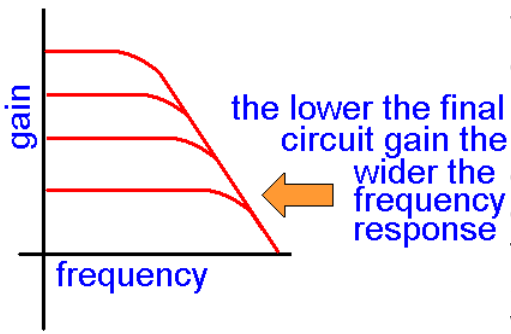
NON-INVERTING GAIN FORMULA



If we assume that an opamp's gain is infinite, or at least several orders of magnitude greater than the final circuit gain, then the voltage gain of a non-inverting circuit of this type is:

$$G_V = 1 + R_1 / R_2$$

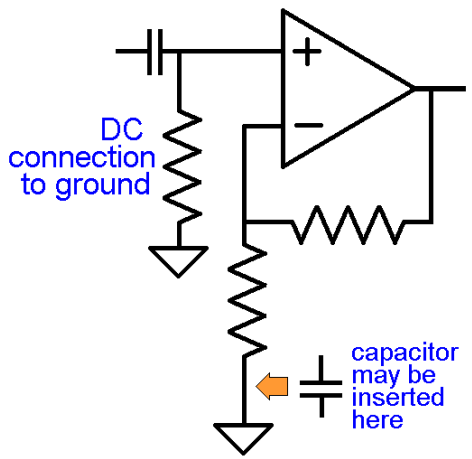
The lower the circuit gain, the wider the bandwidth. This is due to the fact that the opamp's gain falls at high frequencies. Suppose the gain falls to 1,000X at 10kHz.



The final circuit can't have a gain higher than that at 10kHz, because the opamp itself doesn't have that much gain. But if we set the circuit gain to 200X, we can have flat response almost to 50kHz. At a gain of 100, it will be flat to 100kHz. At any particular circuit gain, the frequency where the feedback-determined gain starts to fall is set by the op-amp's own gain-vs-frequency curve.

INPUT GROUND PATH

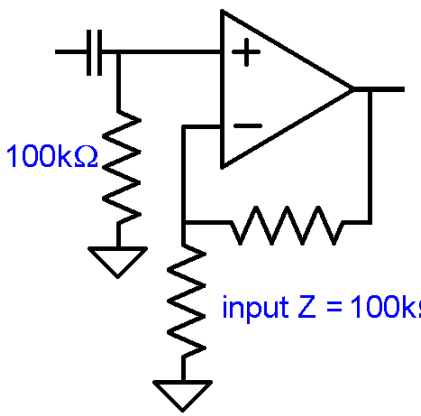
If we don't connect the “+” input to anything, that will result in an uncontrolled and unpredictable DC voltage at that point. We must provide a current path between the “+” input and ground. In the circuits on the previous page, the signal source would have served that function, I was assuming the DC voltage at the signal source's output to be zero. But if there's a capacitor between the source and opamp, we'll need to add a resistor between the input and ground.



The “-” input is usually DC-connected to ground via the feedback voltage divider, but it doesn't have to be; we could break the shunt part of the circuit just above the ground and put a capacitor in. This is often done to achieve a precise DC operating point, as described later, or for other reasons discussed in later chapters.

NONINVERTING AMPLIFIER INPUT IMPEDANCE

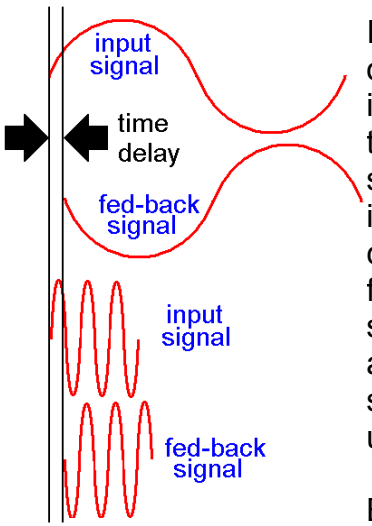
With any decent opamp in non-inverting configuration, where the signal is applied to the “+” input, the input impedance will be very high. An opamp using BJTs for the input differential pair usually has an input impedance of hundreds of kilohms or more. Take, for example, the common “741” opamp, first manufactured in 1968. It's so well designed and so stable in operation that it's still very widely used, almost half a century after its debut. (I bought one at Radio Shack yesterday.) The input impedance is around 2 megohms. The input impedance of the circuit is the parallel combination of this impedance and the input shunt resistor. In the circuit at left, that works out to about 95kΩ.



I mentioned in chapter 3 that many opamps have JFET inputs. A FET-input opamp has essentially infinite input resistance, and only a few pF of capacitance. In most cases, the input impedance will be essentially equal to the resistor between the input and ground. A good one for experimenters is the TL081, or its dual and quad opamp versions, the TL082 and TL084.

HIGH FREQUENCY RESPONSE AND STABILITY

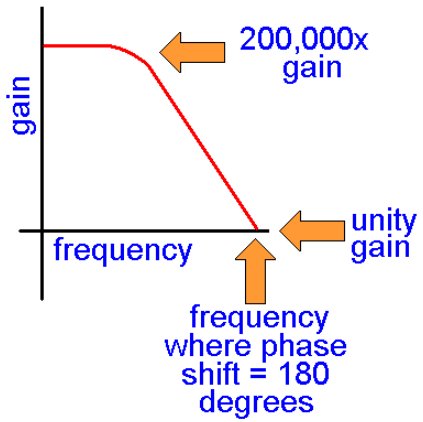
The gain figures mentioned above refer to DC and low-frequency gain. At higher frequencies, the gain falls off sharply, for the same reason it falls off in discrete-transistor circuits: the unavoidable capacitances in the circuit, e.g. Miller capacitances, start shunting the signal.



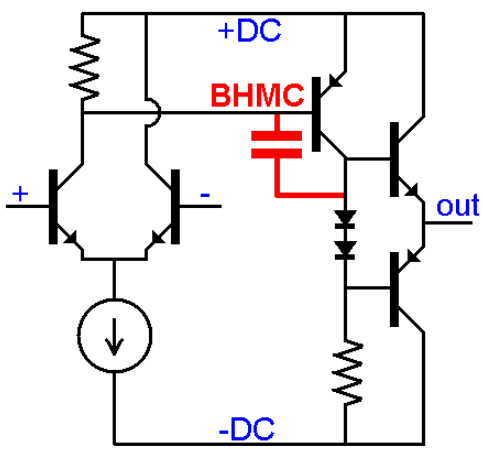
In fact, in a multistage amplifier like an opamp, high-frequency gain *must* fall if the circuit is to remain stable. It takes a small but significant time for a signal to go from the opamp's input to its output, as shown at right. That's called **propagation delay**. At low frequencies, this doesn't matter; the signal fed back to the "-" input is almost completely subtracted from the signal at the "+" input as shown in the upper pair of waveforms.

But at some high frequency, the time delay creates 180° of phase shift so that, as you can see from the lower pair of waveforms, the fed-back signal is added to the input signal instead of being subtracted from it. It's no longer negative feedback but positive feedback.

So, here we are at the point where I left you hanging in chapter 3. An opamp is a multi-stage amp. To recap, when the signal goes through several amplifier stages, phase shift at high frequencies can increase until, at some frequency, the output is 180° out of phase with the "+" input, which means it's *in* phase with the "-" input, meaning negative feedback becomes positive feedback and the amp goes crazy.



By the time that frequency is reached, gain must fall to less than unity. The way an opamp designer does that is to choose one transistor and put a BHMC between collector and base.



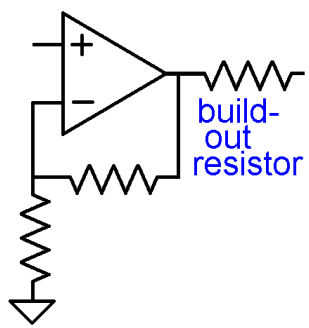
(BHMC is my very own highly technical term meaning Big Honkin' Miller Capacitor.)

So, doesn't that screw up the gain at high frequencies? I'm afraid it does, that's the whole point. The 741's gain starts to roll off at 10Hz. Then again, its starting gain is so high that it doesn't fall to unity until around 1MHz. There are other opamps that still have a lot of gain up into the hundreds of megahertz.

An opamp that is stable in a voltage-follower circuit is called a **unity-gain stable** opamp. If we want to live dangerously, we can back off the Miller cap and make an opamp that is stable at any gain above, say, 10. That opamp will have better high frequency response, and as long as we don't try to put it in a circuit with a gain less than 10, we'll be okay.

It's usually a bad idea to connect a capacitor directly between the output and ground, or to permit a significant capacitance to be connected – a long coax cable, for example. For an amplifier, working into a shunt capacitance is like trying to dance in a bathtub full of axle grease. The capacitor can slow things down to the point where the zero-degree frequency falls well within the range where the opamp still has significant gain, leading to instability.

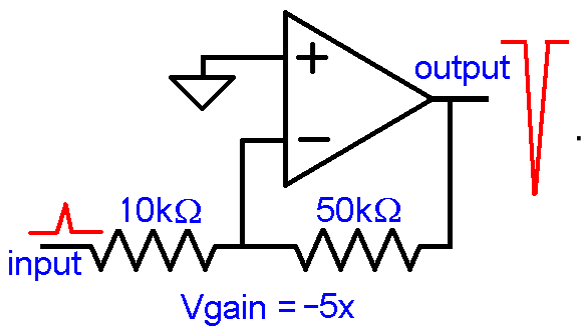
In circuit diagrams, you will often see a resistor in series with the output of an opamp. This is called a **build-out resistor** and its purpose is to isolate the opamp from a capacitive load.



Another purpose for a build-out resistor is to achieve a desired output impedance. Since the opamp has essentially zero output impedance, the final output impedance is the value of the build-out resistor. A video amp, for example, usually consists nowadays of an opamp with a 75Ω build-out resistor.

THE INVERTING AMPLIFIER

Up to this point, we've only looked at circuits where signal is fed into the "+" input. If we ground the "+" input and apply the signal to the "-" input via the feedback network, we'll get an inverted output. This circuit works like the virtual-ground transistor circuit we studied in the last chapter, except that the numbers are more precise and the input impedance is nearer to zero ohms because the active device gain is so much higher.



The gain of an inverting amplifier circuit is:

$$G_V = -(R_1 / R_2)$$

Note the minus sign, showing that the output is opposite polarity from the input.

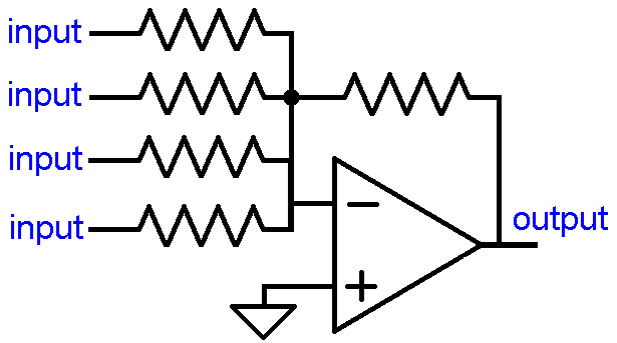
- As before, in normal operation the voltage difference across the inputs is about zero.
- When an input signal causes the voltage at the "-" to change, the opamp responds by creating an output voltage of the opposite polarity.
- This output voltage, via the feedback circuit, drives the "-" back to approximately zero volts.

INVERTING AMPLIFIER INPUT IMPEDANCE

In an inverting circuit, input impedance is not usually as high as in a non-inverting circuit. We discussed the virtual-ground amplifier in chapter 3, but with the opamp's high gain, the impedance at the "-" input really is very nearly zero ohms. Thus the right end of R_2 is essentially shorted to ground, and and $Z_{IN} = R_2$. In the upper circuit at left, the input impedance will be almost exactly 10kΩ.

THE SUMMING AMPLIFIER

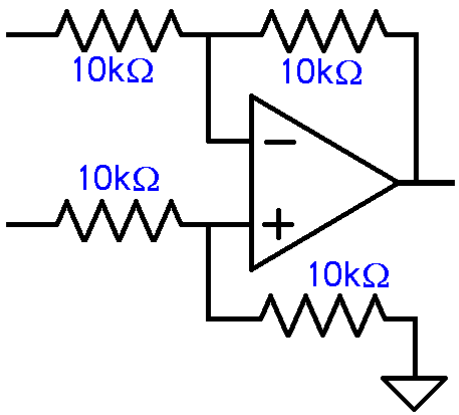
The fact that the "-" input is a virtual ground lets us build a circuit with many inputs. It's just a matter of adding resistors to the input. Note that I've flipped the opamp symbol over so the inverting input is on top; it's just easier to draw the circuit this way.



Each signal source supplies a bit of current through its associated resistor, and the opamp supplies a complementary current to keep the "-" input at zero volts. The output voltage is the sum of all the input voltages times the gain, which is set by the ratio of feedback to input resistors.

One of the virtues of this circuit is that the various signals do not back-feed into each other, because the common summing point is a virtual ground. Another is that a signal can be disconnected at any time without affecting the other signals. The gain can even be made different for each input by choosing different input resistors.

THE DIFFERENTIAL AMPLIFIER



The presence of a differential input allows us to make an amp with a differential or **balanced** input. A balanced circuit is like a differential opamp input; it's "blind" to any voltage that is the same on both inputs. This is called a **common-mode** voltage. Most of the noise and interference picked up by a balanced cable or transmission line is common-mode, and a balanced input rejects it..

With one opamp and four resistors, we can make a simple differential amp. A voltage applied to the "+" input of the circuit is attenuated by 2:1 before it gets to the "+" input of the opamp. If the same voltage is also applied to the "-" input of the circuit, the opamp will, as usual, do whatever is necessary to drive the voltage across its inputs to zero. And, as you will see if you take a pencil and trace voltages, what is necessary is for the output to sit quietly at zero volts. Ergo: common-mode input voltage is 100% rejected.

Or, maybe, 95% rejected if the resistors aren't perfectly matched – which they won't be if we're using standard 5% resistors. Still, 95% rejection is worthwhile; that's 26dB noise reduction, and if the noise is almost acceptable to start with, attenuating it by 26dB will reduce it to insignificance. If we want better, all we have to do is use better-matched resistors.

This isn't the best differential amp. For one thing, the input impedances at the "+" and "-" are equal from the standpoint of common-mode signals, but very unequal from the standpoint of differential-mode signals. In the next chapter I'll show you an improved diff amp using three opamps.

OPAMP OUTPUT IMPEDANCE

The output impedance of a "naked" opamp can be anything, but in a circuit with negative feedback, it's very low. Since voltage gain is governed by the feedback network, output voltage variations caused by load variations will be immediately corrected. In other words, if the opamp's output impedance is not naturally low, it will be driven low by feedback action. In most applications, it is customary to think of the output impedance as being zero ohms.

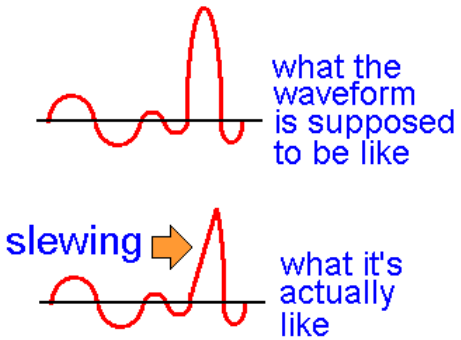
This does not mean you can expect the amplifier to perform correctly if you actually load the output with a very low-impedance load. An opamp has maximum positive and negative output current limitations, which are listed in the published specification sheets. Most have sophisticated circuitry to limit the current so even a dead short won't harm the amp.

SLEW RATE

An amplifier's response to a voltage change at the inputs is not instantaneous, and that applies to opamps as well.
If the voltage difference at the inputs tells the output to go positive by 1V, it will do so – but it will take a little time to get there. "Little time" may be very little, nanoseconds for a fast amp, but it's not zero.

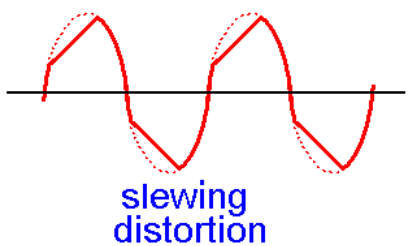
When an amp is used well within its limits, the output will follow the input in whatever way the feedback network dictates. But if the signal demands a greater rate of change than the opamp is capable of, it starts to **slew**, meaning that it rises or falls at its maximum speed, trying to catch up to where it should be.

The illustration above right represents an opamp that properly follows the signal most of the time, but when the signal suddenly goes way positive, it can't keep up. It goes in the direction it knows it's supposed to go, as fast as it can, until it reaches where it should be, and then it starts behaving correctly again.



Slew rate is typically measured in volts per microsecond. For example, the 741's slew rate is a bit under 1V/µs. That means, if the output starts at 0V and we make it rise for 3µs, it will have reached a voltage of a bit under 3V.

The voltage differential across the inputs doesn't matter; the output can only rise or fall so fast, and that's it. We can't make it rise faster by driving the input harder.



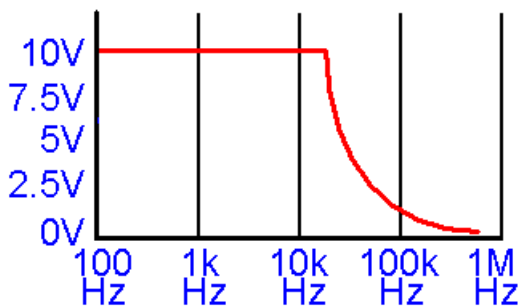
Slew rate sets an absolute limit to the highest undistorted sinewave output voltage an opamp can achieve at a specific frequency; the higher the frequency, the lower the maximum voltage.

Below that limit, negative feedback works fast enough to control the output, and the output voltage will be a faithful copy of the input voltage. Above the limit, the opamp won't be able to keep up, and the waveform will be distorted.

MAXIMUM OUTPUT VOLTAGE

The maximum instantaneous output voltages in either direction are dictated by the DC supply rail voltages. The output can never go higher than the positive rail or lower than the negative rail. Many opamps don't even get close to the rails. Those that do are advertised as **rail to rail** opamps. That's important in low-voltage applications.

Manufacturers publish graphs that show maximum output voltage versus frequency. These are often called large signal graphs. At low and moderate frequencies, maximum output is limited by supply rail voltage. At some high frequency, slew rate comes into play, and then maximum output falls by half for every doubling of frequency.



For the 741, that critical frequency is about 10kHz. Many micropower opamps are much slower; others can do full output to 100MHz. Bottom line: you can't just replace any old opamp with any other old opamp. You need to look at the highest frequency it will have to handle, as well as other specs that may be critical such as maximum output current. And if you're putting it in a unity-gain or low-gain circuit, you have to check the specs to be sure the opamp will be stable at that gain.

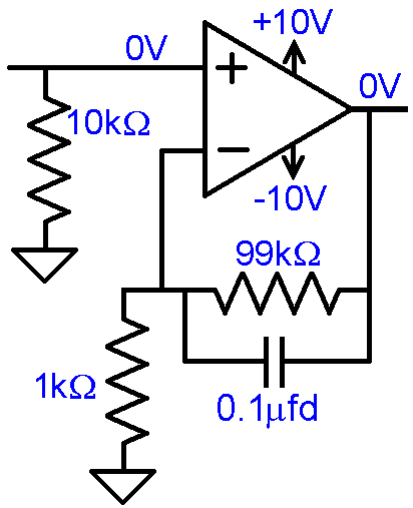
DC AMPLIFIERS

With a bipolar supply, the input and output voltages can be positive, negative or zero. We don't need blocking capacitors to DC-isolate the amplifier. This means an opamp can be a DC amplifier as easily as an AC amplifier.

To return to the example used in the previous chapter, if we have a single thermocouple that puts out only millivolts, an opamp circuit can boost this voltage to the point where it can be used in measurement or control circuits.

There are so many applications for DC opamp circuits in instrumentation and industrial control that I'm not even going to try to list them. However, it's worth asking the question: what's a DC “signal?”

Usually, it means a voltage that changes so slowly, like the water level in a tank, that we look at it on a moment-to-moment basis, without thinking about an overall waveform. Very often, the feedback network will deliberately limit high frequencies to discriminate against noise. In the amp at right, low frequency gain is 100X, but response starts to fall off at 16Hz.



Since we're dealing with voltages that change slowly, we can get immense amplification if we need it. Even the ancient 741 can achieve DC gain of 50,000 with room to spare.

Because an opamp used for a DC amplifier doesn't need to have good high frequency response, it can operate on very little power; a lot of current doesn't have to be available to quickly charge and discharge those Miller capacitors. Manufacturers have been busy of late, coming up with lots of opamps that draw only microamps of current. The advantage is not in energy savings as such – normal opamps draw only a few milliamps – but in extending battery life in wireless applications.

OFFSET VOLTAGE

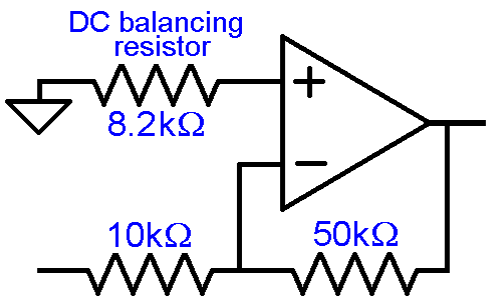
I've been ignoring a monkeywrench, or at least a fly in the ointment, that crops up whenever an opamp is part of a DC measuring circuit. You know that both BJT and FET characteristics vary all over the map. But when they're manufactured on the same substrate, within thousandths of an inch of each other, they vary *together*. If one transistor of a diff pair has abnormally high gain, so will the other. They're automatically well matched.

But they're never *perfectly* matched. There's always a few mV of **offset** between them. The input differential that results in a zero-volt output is not zero volts, but a mV or two one way or the other.

In an AC amplifier, this doesn't matter, because there's usually a blocking capacitor in the shunt leg of the feedback circuit. That means the DC gain is unity, so 2mV of offset at the input results in 2mV of offset at the output. Not a big problem.

But with a DC amp, we want DC gain, so we can't put that capacitor in. Offset will then be amplified by the same factor as signal. If we set up the circuit for 1,000X gain, a 2mV offset at the input will become a 2V offset at the output.

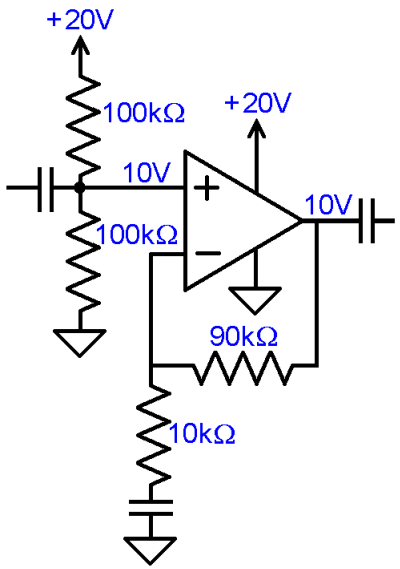
The way we handle this is to **null out** the offset with a trim pot. An opamp may have nulling terminals. If it doesn't, we can inject a small correction current into the "-" input. These methods are beyond the scope of this tutorial, partly because different opamps use different methods. But the subject is covered in downloadable application guides offered for free by IC manufacturers.



If the input transistors are BJTs, as they usually are, input current will be low but not zero. This is a further source of error if the circuit resistance to ground is different between the "+" and "-" inputs. The error can be reduced by adding a resistor between the "+" input and ground to create a voltage drop equal to the drop across the feedback resistors.

SINGLE-SUPPLY OPAMP CIRCUITS

In an AC circuit, an opamp's power supply doesn't have to be bipolar. If the circuit is set up so its quiescent operating voltage is, say, +10V instead of zero volts, it'll work fine.



For AC amplifiers, you just need capacitors at the input, output and feedback circuit, as shown at left. DC circuits are a little more complex; the DC voltage you're measuring must be referenced to the same +10V that the opamp circuit is referenced to. It is often easier to derive an artificial ground from a single supply, which is something you may run into in equipment you're working on. I'll describe that in chapter 9, which deals with power supplies.

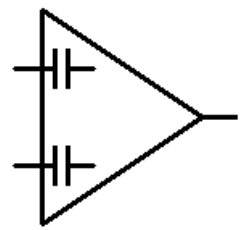
HOW NOT TO DESTROY AN OPAMP

Opamps are not hard to break, though they're not particularly delicate if used correctly. The following rules must be observed.

- Never reverse the polarity of the DC power supply terminals. That is, the positive supply terminal must never be connected to a voltage lower than the negative supply terminal, even for an instant.

- Never exceed the maximum voltage across the DC terminals specified by the manufacturer. For most opamps, this is 36V, or + and – 18V with a bipolar supply. Some special opamps handle higher voltages, while many video and low-power opamps have maximum power supply voltages of 5V to 12V.
- Never drive any terminal more positive than the positive DC supply terminal or more negative than the negative DC supply terminal. The DC rails should be considered absolute voltage limits for every terminal on the IC. In particular, never connect the inputs to a signal source that may exceed the supply rail voltages. I said at the beginning that the absolute input voltages are irrelevant, but that's only true so long as they stay between the supply rail voltages. If either the "+" or "-" input is driven hard to a voltage outside the supply rails, the opamp may be ruined.
- Never connect a capacitor across the output and ground. Exception: some special ICs are designed to drive capacitive loads. If you run into a piece of equipment where this is done, and it doesn't cause problems, the opamp is probably one of those special ones.
- Never jam a signal backward into an opamp output. Protection against short circuits and protection against a backfed signal are two different things. In particular, this means do not parallel two opamp outputs, as each will try to ram its own signal backward into the other. If you must parallel outputs, use build-out resistors to give them a bit of isolation from each other.
- It's good practice to bypass the DC rail or rails to ground with a capacitor near the IC. A common value is 0.1µfd. This helps prevent instability and oscillation.

There's one other issue I need to mention. Some special-purpose opamps have CMOS inputs. These may or may not have good protection; too much protection screws up the extremely high impedance of the inputs, which is the main reason for CMOS opamps to exist. As you gain experience, you will get to know which equipment has CMOS opamps. Unfortunately, there is usually no call-out in schematic diagrams or on the PC board next to the opamp. If you're very lucky, you'll see a symbol something like this. (There doesn't seem to be a standard symbol.)



MOSFET-input opamps should be treated with the same care as small MOSFETS: keep the opamp in a conductive bag until use, ground the soldering iron, the benchtop and yourself through 1-meg resistors, etc. When you're soldering one in, connect the power supply pins first.

End of EM3 Chapter 4

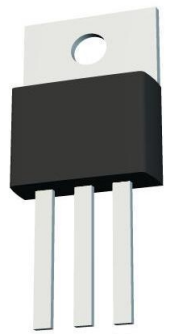
EM3 Chapter 5
ANALOG INTEGRATED CIRCUITS

Objectives of this chapter:

- *define "discrete" and "integrated" as they apply to semiconductors*
- *explain advantages of integration*
- *define and discuss "comparator"*
- *discuss analog switches and why they are used*
- *describe the operation of electronic potentiometers*
- *list the requirements of a video amplifier IC*
- *define and discuss "H-bridge"*
- *describe the operation of a 555 timing IC*
- *explain the functinon of a LED driver IC*
- *explain the function of a Hall Effect sensor*
- *define "MEMS"*
- *describe the input/output characteristic of a logarithmic amp*
- *define "VCO"*
- *discuss phase locked loops*
- *define "ASIC"*
- *explain how IC pins are numbered*

DISCRETE AND INTEGRATED DEVICES

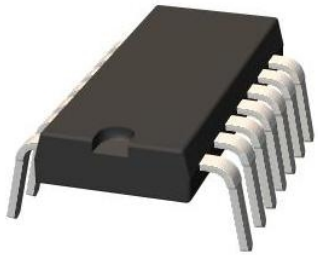
In the last chapter, I sort of plunged into integrated opamps, because it was a logical progression from transistor amps. I want to begin this chapter by backing up and looking at the significance of integration as such.



As mentioned previously, an individual active device, e.g. a single transistor, is called a **discrete** device. ("Discreet" means something different. If you used "discreet" to refer to a discrete device, I would be discreet enough not to laugh.) If you want to build a piece of electronic equipment the discrete way, you gather several transistors, a bunch of resistors and other components, and connect them all together.

We've seen that discrete transistors, diodes, and other solid state devices are "grown" in batches on a substrate. Back in the 1950s, it occurred to someone that, if they didn't cut all the devices apart, leaving substrate chips with several devices each – and if they connected the devices – if they added patterned layers of conductors and resistors – it would be possible to create complete circuits instead of individual devices.

That's what an **integrated circuit** or **IC** is – many components in one package, requiring few external parts to make a complete circuit.



The word **chip** is often used to informally describe an IC, i.e. a small chip off a silicon substrate disc. Most of the IC's we will deal with come complete with packages and leads. IC chips without packaging or leads are also sold for imbedding into mass-produced products. Once installed, a leadless chip is usually covered with a blob of hard black plastic. This makes the IC inaccessible; if it dies, you replace the whole circuit.

One of the great virtues of IC's is that they combine many parts in a single package. That reduces the cost of parts, but a far greater benefit is that it reduces the cost of manufacturing a piece of equipment. Discrete circuits have, in addition to transistors, lots of resistors, capacitors, and occasionally inductors and other components. In a manufacturing situation, each of these components must be individually specified, ordered, paid for, inventoried, tested and assembled on a PC board; then the whole circuit must be tested again. If a single part is back-ordered, the whole production line grinds to a halt. Integrated circuits, with many parts contained in a single package, make this job a lot simpler and less expensive, and usually improve performance to boot.

Another virtue is that the all the transistors are almost perfectly matched. Individually, they vary as much as discrete transistors. But being manufactured together as they are, they vary together. They can hardly help but have the same gain and base-emitter dropping voltage, two critical characteristics.

A third virtue is reliability. In a discrete circuit there may be dozens of parts, and if any one of them fails, the whole unit fails. The failure rate of ICs is similar to the failure rate of any *one* component in a discrete circuit, which means the overall failure rate in a complex system goes way down. This is immensely important in computers, which contain ICs with billions of transistors.

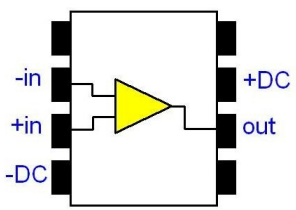
It used to be that IC's were limited to low frequencies, leaving high frequencies as the sole preserve of discrete amplifiers. However, IC's are now available that work at such high frequencies that discrete transistors have become obsolete in many applications.

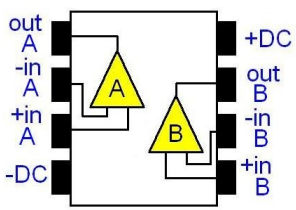
Analog ICs come in so many varieties that, if I tried to describe all of them, there wouldn't be room to give you useful information on any of them. So, beginning on the next page, I'm just going to describe the most common types.

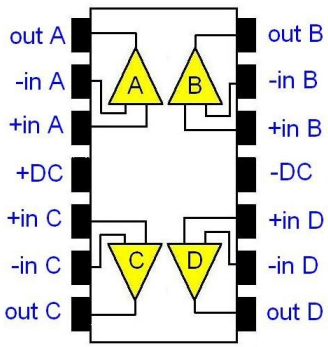
When you find an IC in a piece of equipment you're working on, it will almost always have a part number stamped on it and you can usually go to the web and look it up. I say “usually” because some part numbers are proprietary, and the info is unpublished to keep people from stealing designs.

OPAMPS

Probably the most common analog IC is the operational amplifier, which usually comes in 8- or 14-lead packages containing 1, 2 or 4 opamps. There's a more or less standard lead connection pattern, called the **pinout**, for each of these packages.

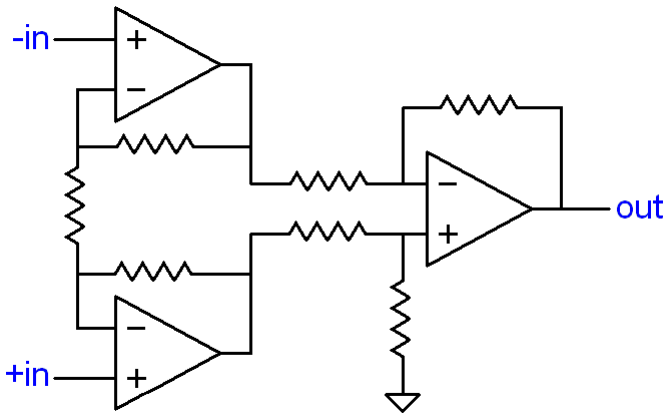






THE INSTRUMENTATION AMPLIFIER

In chapter 4, I mentioned that the 1-opamp differential amplifier has limitations. It's good enough that it's widely used in, for example, the balanced inputs of audio power amps. But it's nowhere good enough for precision instrumentation.



The three-op-amp circuit above is called an **instrumentation amplifier**. The opamp on the right works like the one-opamp differential amp; the two on the left create an interface that presents a well matched, well balanced, high impedance load to the signal source.

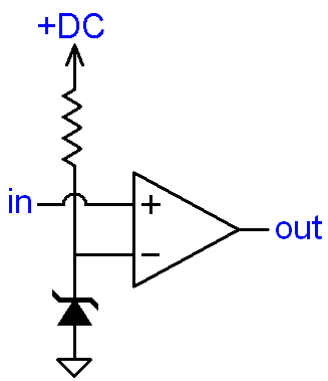
COMPARATORS

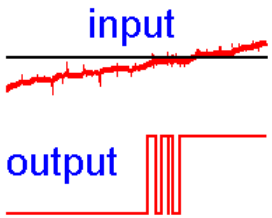
A comparator, as its name implies, is a chip that compares two voltages and indicates which is higher. It is essentially a modified op-amp, and the schematic symbol is the same. It has very high gain and can distinguish between two voltages that are a fraction of a millivolt apart.

A comparator cannot be used as an op-amp because it is not stable in a circuit with negative feedback. The manufacturer removes the Big Honkin' Miller Capacitor mentioned in chapter 4, deliberately giving up stability in return for increased speed. This is no great sacrifice, since a comparator isn't intended for amplifier use.

With an opamp, we usually don't want the output to hit the rails, but a comparator is supposed to hit them. It's always either full “up” or full “down” depending on which input is higher.

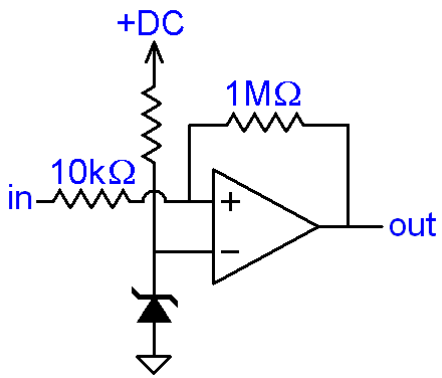
A comparator is most often used to compare an input voltage to a preset reference voltage. If the input voltage is connected to the “+” comparator input, the output will go high when the input voltage is higher than the reference.



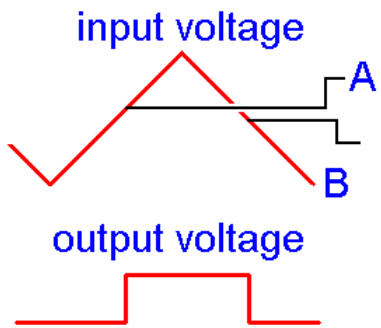


If there is noise on the input voltage, the comparator output may be unstable when the input and reference voltages are close together. The illustration at left shows a rising input voltage with noise. Note that, when it is near the reference (black) voltage, the output goes back and forth several times as the noise obscures the moment when the input voltage passes the reference voltage.

This can be avoided by feeding back a little of the comparator's output to the “+” input. This is the opposite kind of feedback from what we usually want. It's called **positive feedback** and it has the effect of reinforcing any change in the output. In this case, when the output voltage is low, current through the feedback resistor pulls down on the “+” input just a little.



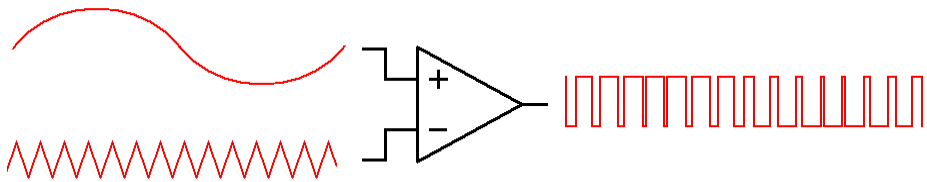
This means the input voltage has to go a little higher than the reference voltage to make the output go high (“A” at right).



Once the output does go high, feedback current pulls up on the “+” input, so the input voltage has to go significantly lower to make the output go low again (“B” at right). In other words, the input voltage that makes the output go high and the input voltage that makes the output go low are not the same voltage. The input voltage has to go a bit higher than the reference voltage to make the output change, and then has to go a bit lower than the reference in the other direction to make the output change again. This is called **hysteresis** (hiss-ter-EE-siss). A comparator with positive feedback is a bit like a snap switch, which has to be pushed beyond the halfway point to turn on and pushed past the halfway point in the other direction to turn off.

A circuit that uses positive feedback to reinforce and speed up a transition is called a **Schmitt trigger** after the late Otto Schmitt, who also invented the opamp.

A simple class-D amplifier can be made from a comparator and a triangle-wave generator. During positive input signal alternations, the “+” input will be higher than the “-” input most of the time, and the output will be high most of the time. During negative input alternations, the “-” input will be higher than the “+” input most of the time and the output will be low most of the time. This circuit is also the basis of many switching-type power supplies.



An opamp works fine as a comparator, as long as it's fast enough to work in the particular application. That's often no challenge at all, because many comparator applications involve very slow signal voltage changes – furnace thermostats, for example. That's also a good example of an application where hysteresis is used. You don't want the furnace cycling on and off every time the temperature changes by 0.1 degree, so you build in some hysteresis.

Comparator pinouts are not as standardized as opamp pinouts; there are many variations. When you see a comparator in a schematic, you have to pay attention to the specific pinout for that part.

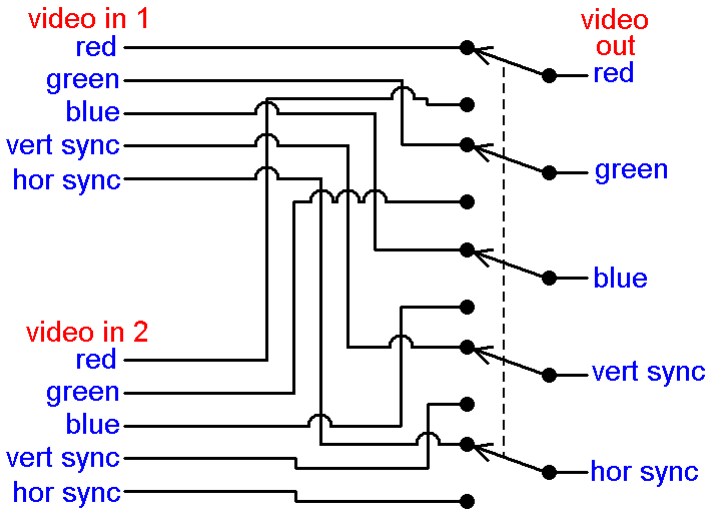
ANALOG SWITCHES

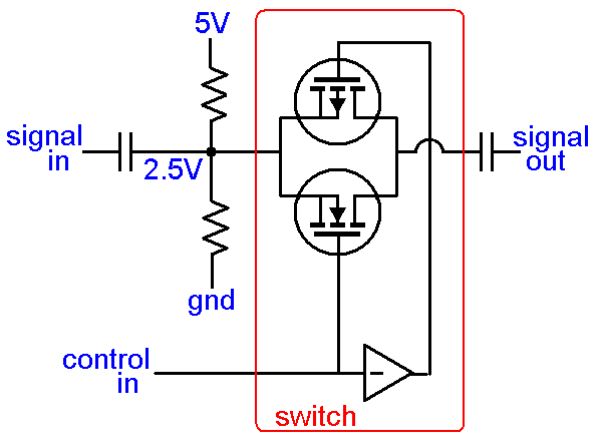
There are a number of good reasons why you might not want to have to use a mechanical switch to switch a signal on and off, or switch between two signals.

- You might want the switch to be electronically controlled, by a computer for example.
- You might be dealing with such high frequencies, or such fragile signals, that you don't want to send them through a switch located some distance from the rest of the circuit, which might degrade the signals.
- You might want to switch a number of signals at once. For example, switching VGA video involves five separate signals: Red, Green, Blue, Horizontal Sync, Vertical Sync.
- The physical assembly and wiring associated with a complex mechanical switch often costs more than doing the same job with an electronic switch.

The diagram above right is a mechanical VGA switch. Aside from the time and cost of assembling this rat's nest of wires, there's a great potential for signal degradation. If all of that can be done inside a chip, we're way ahead.

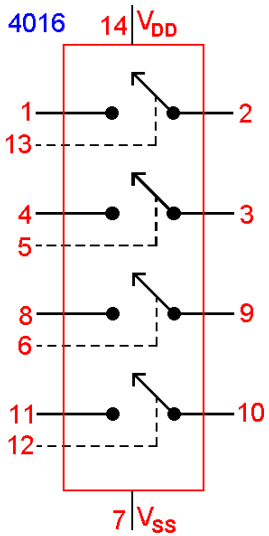
The obvious choice for an electronic switch is a MOSFET. When the channel is in conduction mode, the switch is "closed." When it's in pinch-off, the switch is "open." The signal voltage must be confined between the switch's supply rails, so its DC level is set at half the supply voltage.



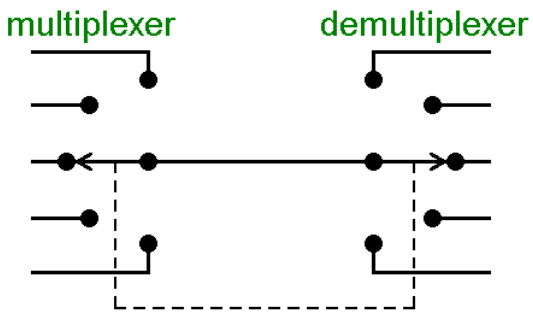


Usually, an N-channel and P-channel FET work together to insure that channel resistance is low over the entire signal voltage range. In the illustration at left, the supply voltage is 5V.

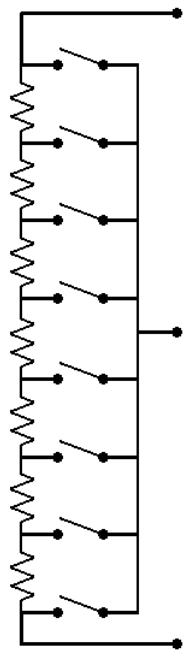
The quiescent voltage of the signal path is set by a divider at 2.5V. Note the inverter which insures that, when a positive voltage is applied to the gate of the NFET, the gate of the PFET goes to ground; thus the two turn on and off together.



The classic analog switch is the 4016 or 4066, a CMOS IC with four separate SPST electronic switches. The $\mathbf{V_{SS}}$ lead is grounded and the $\mathbf{V_{DD}}$ lead is connected to +5V. Though the 4016 and 4066 remain viable and widely used, a great many newer and better switches are now available. Many are designed for specific applications – for example, 5-channel units, essentially 5PST or 5PDT switches in one IC package, especially designed for VGA.



Some IC switches are called **multiplex** IC's. Multiplexing means selecting one signal from a number of signals and connecting it to a single communication channel. At the other end, another switch may **demultiplex** the signals, distributing each to its proper destination. The distributor in an older car engine is a demultiplexer, distributing spark voltage to the plugs. Many signals can be sent simultaneously over one communication channel using a very fast multiplexer on one end and a synchronized fast demultiplexer on the other end. If multiplexing goes through at least two full cycles for every cycle of signal – that is, if the multiplexing cycle frequency is at least twice as fast as the highest frequency contained in any of the signals – all the signals will come out undamaged. This is called **time division multiplexing**. One example is the wired long-distance telephone network, which used to make extensive use of it.



POTENTIOMETERS

For the same reasons that an electronic switch is useful, an electronically controlled potentiometer is often useful. Most electronic pots consist of a series of fixed resistors, with the “wiper” consisting of FET switches and a circuit to choose which switch is closed. This is called a **stepped potentiometer**, and if the steps are small enough, it's almost as smooth as the continuous kind.

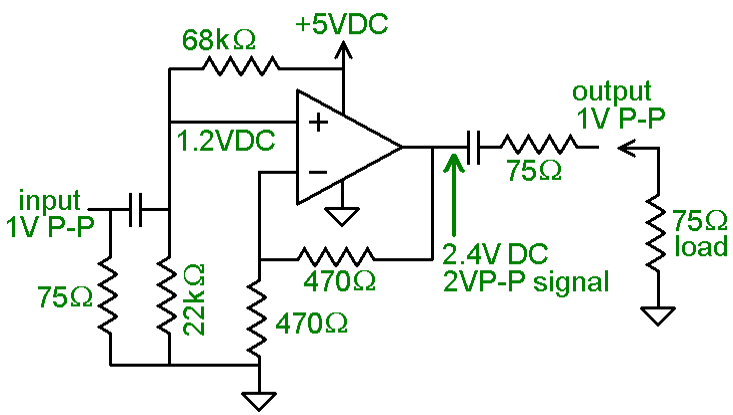
Many IC pots are designed to be controlled by up-down pushbuttons, like the volume controls on a cellphone. Others are controlled by digital codes, for the sake of easy interface with computers.

In audio applications, electronic pots are often incorporated into chips that control such things as bass, treble, and balance. This is especially handy in 5.1 or 7.1 channel applications where a mechanical control would be clumsy.

BASEBAND VIDEO AMPLIFIERS

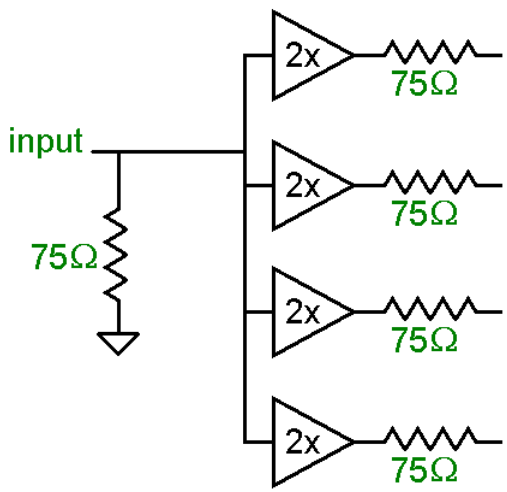
A baseband video amp is an amplifier designed for the specific requirements of video:

- a P-P signal of a few volts at most
- a bandwidth from 30Hz to many MHz
- an output capable of driving a 150Ω load



(If you've forgotten what "baseband" is all about, review chapter 1.)

A baseband video amp, when properly connected to a 75Ω load, is usually designed to have a gain of unity. That's most often achieved by combining a zero-output impedance amp with 2x gain and a 75Ω build-out resistor. This resistor and a 75Ω load form a 2:1 voltage divider, bringing the overall gain back to unity.

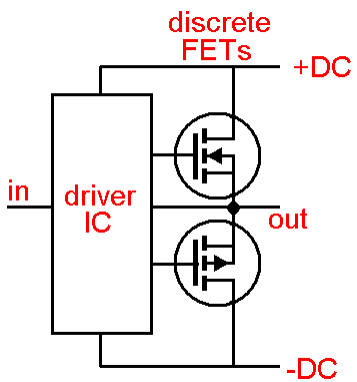


This unity-gain amp is often used as part of a **distribution amplifier**, where one video signal must be routed to multiple destinations without upsetting levels or impedances. Since each amplifier circuit has a high input impedance, one 75-ohm resistor can serve as the input termination for all of them.

Some video amps have extra functions associated with maintaining a specified DC level, which is often critical in video circuits.

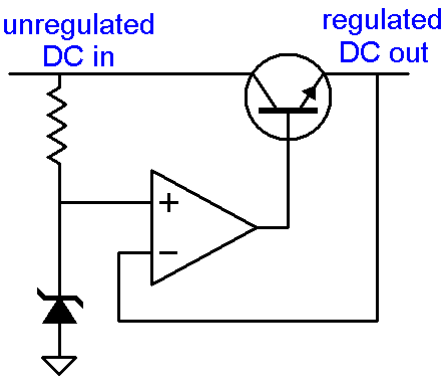
DRIVER IC'S

High-power circuits involve voltages and currents that exceed the capabilities of IC's. For example, the output circuit of a large audio power amplifier or DC motor controller might have to dissipate hundreds of watts, which would melt an IC in no time. In such a case, an IC might be designed to drive and control external transistors that do the actual heavy lifting. Integrating the low-power part of the circuit simplifies construction and also lets us include sophisticated biasing and protection circuitry.



VOLTAGE REGULATION

If we take an amplifier with a high-current output and connect its input to a stable reference voltage, we have a linear voltage regulator which can turn unregulated DC with a lot of ripple into a smooth, ripple-free, stable DC voltage.
In the circuit at right, note that the transistor is in common-collector mode, meaning there is no inversion between base and emitter. The opamp and transistor together, with the output connected back to the "-" input, form a high-current voltage follower. The stable zener voltage, connected to the "+" input, is duplicated at the output.



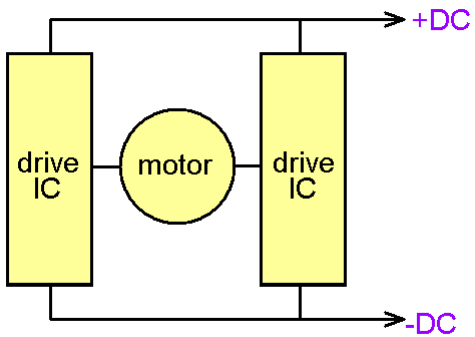
So much electronic equipment requires regulated DC that IC manufacturers have responded with hundreds of kinds of integrated regulator chips.

Where the power dissipation involved in regulating high-current DC would create too much heat, class D switching amplifier techniques can be used to regulate a voltage without burning up a lot of power.

Regulators amount to only a small portion of the chips available to perform various power-supply related functions. We'll look at more of them in chapter 9, which deals with power supplies.

MOTOR DRIVES

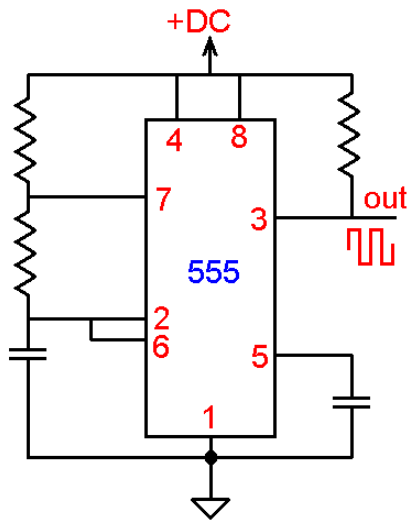
The speed of a DC motor depends on voltage. Of course, it's possible to control a voltage using big transistors and simply dropping a high supply voltage down to whatever you want to drive the motor with. But this is inefficient, involves dissipating a lot of power, and is completely impractical with multiple-horsepower motors. However, the same thing can be done with very little wasted power by using switching, or class D, technology.



Usually, the motor will have to be able to turn either way, meaning the voltage across it may have either polarity at a given moment. To allow that to happen, the motor is connected *between* two push-pull pairs of transistors. This is known as an **H-bridge**. And, if we want to control speed, a pulse-width circuit connected to an H-bridge is the easiest way to do it.

AC motor drives have to do more than adjust voltage. The speed of an AC motor depends almost entirely on the AC frequency. Motors that run on 60Hz power naturally turn at 3,600, 1,800 or 1,200 RPM, minus a few percent for "slip." Getting them to run efficiently at other speeds involves conversion of 60Hz power to some other frequency. This involves so much complex circuitry that it would be completely impractical without circuit integration. We'll be looking at this in greater detail in EM5.

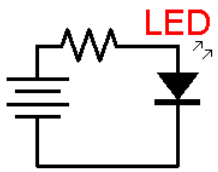
TIMING CIRCUITS



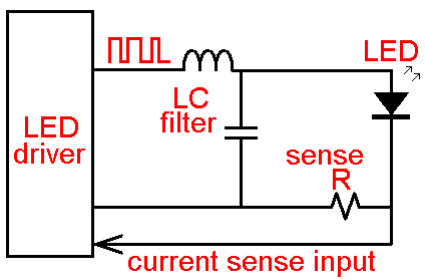
A timing circuit creates a pulse of a certain length, or after a certain delay. Op-amps and comparators can be used to make timing circuits, and often are, but there are also dedicated IC's available for the purpose. By far the most common is the "555" IC, which has been the standard timing chip for decades and can make an oscillator, pulse-width modulator and other things as well. The circuit at right shows a 555 set up as a square or rectangular wave generator.

LED DRIVERS

A light emitting diode is like any other diode in one respect: it isn't easy to control current through it by controlling voltage across it. For this reason, current has traditionally been controlled by putting a dropping resistor between the DC source and the LED.

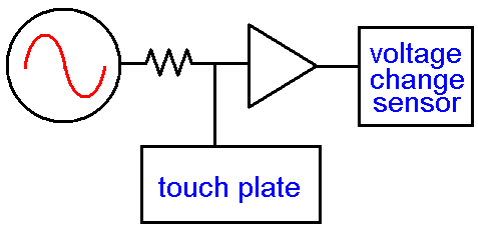


This works fine with standard low current (e.g. 30mA) LEDs, but now that we have LEDs that draw half an amp, the wasted power has become an issue. It is possible to eliminate most of the waste by using switching technology which is closely related to class-D amplifier technology. A high-power LED driver begins with a solid-state SPDT switch that cycles between +DC and ground at perhaps 50kHz. An LC filter between the switch and the LED turns the chopped DC into a continuous current. We've seen this before; the added element here is a resistor that develops a small voltage drop. The LED driver measures this drop to tell how much current is flowing in the LED, and adjusts the duty cycle of its rectangular-wave output accordingly.



TOUCH CONTROL IC'S

Probably the oldest touch-controlled circuit is the table lamp with no switch; you touch the lamp's metal base to turn the light on and off. This circuit takes advantage of the fact that a human is a big bag of salt water, and has significant capacitance. The easiest way to sense touch is to detect the 60Hz energy that a person picks up from the environment. A more sophisticated method is to generate an AC waveform, feed it through a high resistance, and detect a change in loading when a finger touches the plate. (The resistance and your body capacitance act as a voltage divider.) The modern descendents of this simple circuit are found in everything from appliances to smartphone screens.



PROXIMITY SENSORS use a similar technology, but are set to detect the subtle change in the ambient capacitive field that occurs when a person gets close.

HALL EFFECT IC'S

Current flow through a Hall effect sensor is proportional to the magnetic field surrounding it. In a way, it resembles a DC transformer. It can be used with a magnet as a proximity or position sensor. Some Hall sensors' outputs are linear and proportional; others incorporate comparators, changing them from linear sensors to threshold sensors that act like switches.

ACCELEROMETERS AND POSITION SENSORS

These IC's use **MEMS** technology, which stands for **microelectromechanical system**. (Try spelling that from memory.) MEMS IC's have microscopic machines built right into the silicon structure of the chips. These machines can be made to sense orientation and movement. A MEMS chip is what enables a smartphone to reorient the display when you rotate the phone.

Speaking of smartphones, a lot of phones now use MEMS microphones, where a microphone, amplifier and analog-to-digital converter are integrated in a single chip. You're going to see a lot more MEMS technology in the future, because the ability to “print” tiny machines on a silicon wafer the way transistors and IC's are printed, thousands at a time, offers enormous potential to reduce size and cost.

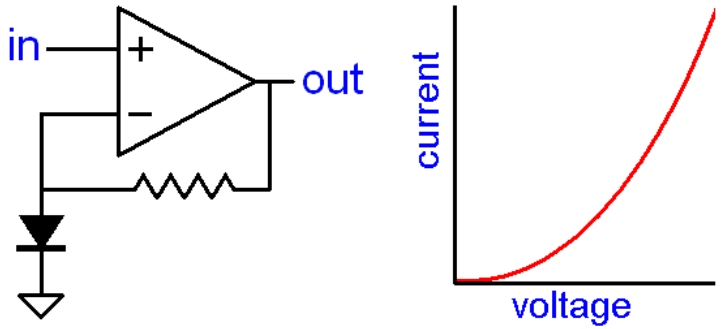
TEMPERATURE SENSORS

I've mentioned thermocouples and thermistors, but the 0.7V forward drop across a diode can be used as a temperature sensor too. This drop falls a tiny bit with rising temperature, about -2mV per degree C or a little over -1mV per degree F. This change, amplified and compared to a known reference voltage, can be used as a thermometer.

LOGARITHMIC AMPLIFIERS

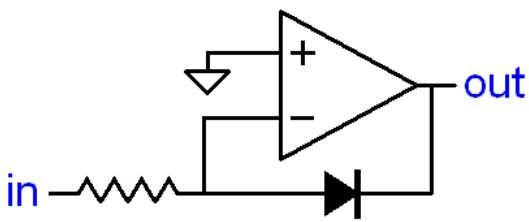
Most of the amplifiers we've studied to this point are, ideally, linear. A log amp, on the other hand, has a logarithmic input-to-output transfer function. When a log amp is placed in a feedback circuit, the opposite (antilog) characteristic is obtained. Together, a log amp and an antilog amp are the building blocks of an RMS voltage or average power meter.

Our rule of thumb – a silicon diode's forward voltage drop is about 0.7V – is just an approximation; it actually varies with current a little. It so happens that a diode junction's voltage vs. current curve is logarithmic. For silicon, every 10x multiplication of current results in about 0.1V rise in forward voltage. This is a problem when we want to build a linear amp, and it's why all that negative feedback stuff is necessary. But it's cool when we want to build a logarithmic amp.



The simplified log amp above won't actually work satisfactorily; there are all sorts of biasing and temperature stability issues. But it shows the concept. If the voltage at the "+" input varies linearly, the same voltage will appear across the diode. But, since it takes a logarithmic change in current to get a linear change in voltage across a diode, and that current comes from the output via the feedback resistor, the output voltage will be a logarithmic function of the input voltage.

If we move the diode from the shunt to the series position, and feed a variable voltage into the "-" input, we get the opposite characteristic, an antilog amplifier.

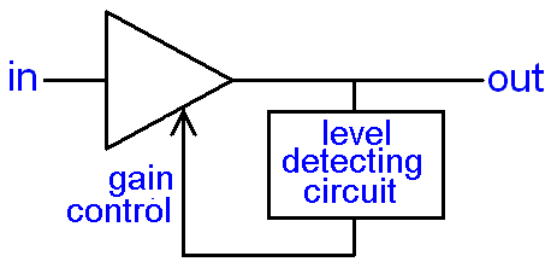


VARIABLE GAIN AMPLIFIERS

The classic variable gain control is, of course, the potentiometer, but that's a mechanical part. It's often necessary to do that job electronically. The one I showed you above, whose gain changes in steps, is a combination of opamp and stepped resistor string. There are others which change gain smoothly, using the logarithmic characteristic of base-emitter junctions to vary gain. This, of course, results in signal distortion. There are clever circuits that apply another junction's logarithmic characteristic to the distorted signal in the opposite way to clean it up again. "Variable gain amp" is often shortened to **VGA**, not to be confused with the computer display wiring standard.

AUTOMATIC GAIN CONTROL

If we connect a variable gain amp to an automatic level detection circuit, we get an automatic gain control or **AGC** function.



One example is found in the input circuitry of an AM radio. It is unacceptable for a 50kW station to come through 10dB louder than a 5kW station. You expect all stations to be about the same loudness, meaning the gain must change when you change stations. Back in the 1920s, you had to make this adjustment manually. But since the 1930s, radios have incorporated AGC circuitry to automatically correct for varying signal strength. In EM6, which deals with RF and wireless equipment, we'll be looking at AGC in detail.

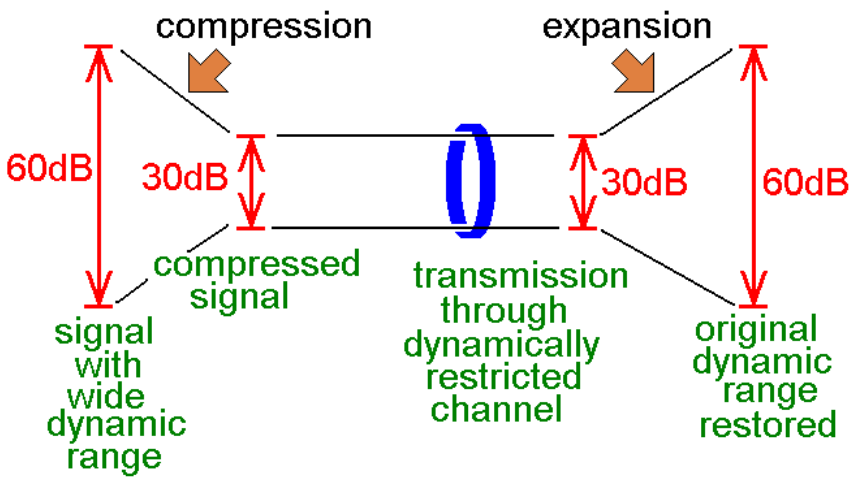
COMPRESSION AND EXPANSION

AGC isn't just used to adjust a radio station's signal strength; it's applied to the program material too, at the radio station, to reduce loudness variations. When AGC is used to reduce the dynamic range of a program signal, e.g. reducing a 20dB loudness variation to 10dB, it's called **compression**. If you've listened to a station playing music with a heavy beat, and noticed that every drumbeat punches a hole in the vocals, that's compression. Every time that big drumbeat comes along, the AGC turns the gain down, then turns it up again.

Note that, in digital technology, “compression” means something different. Digital compression means reducing the amount of data. Analog compression means reducing dynamic range.

When AGC is used to *increase* the dynamic range of a signal, e.g. turning a 10dB change into a 20dB change, it's called **expansion**. The primary use for expansion is to reverse compression applied earlier in the signal chain.

If a transmission channel has limited dynamic range, the signal can be compressed before transmission and then expanded upon reception, making it possible to transmit information that has a wider dynamic range than the channel it is transmitted through.



Complementary compression and expansion is called **compansion**. Telephone companies use this a lot. Optical film sound tracks, in the absence of compansion, sound really rotten due to their limited dynamic range. One of the things that first put the Dolby corporation on the map was a sophisticated compansion system that makes optical sound tracks sound good. You'd be surprised how often you're listening to a Dolby-ized analog sound track in a theater, because the digital sound track has been damaged.

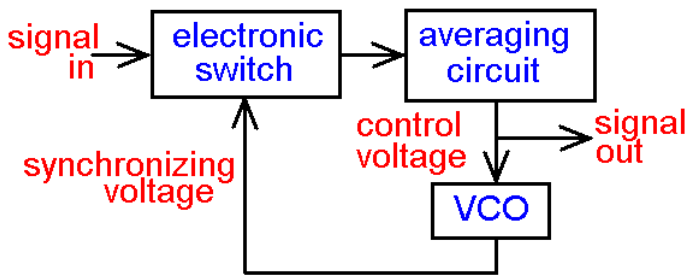
VOLTAGE-CONTROLLED OSCILLATORS

Quick review: an **oscillator** is a circuit that generates a continuous signal at a particular frequency. A voltage-controlled oscillator or **VCO** is an oscillator whose frequency is determined by a control voltage. A 555 is one such oscillator. Refer to the chapter 6 for a discussion of the 555.

THE PHASE LOCKED LOOP

A phase-locked loop or **PLL** is based on a VCO connected to an electronic switch. An input signal is fed into the switch; the output voltage of the switch is a voltage whose instantaneous value corresponds to the amplitude, frequency or phase of the input signal depending on how the PLL circuit is designed.

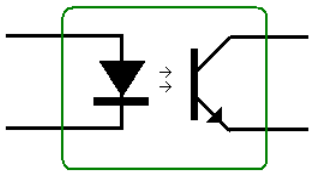
This output is fed back to the VCO to control its frequency; thus the VCO frequency locks itself to the input signal frequency. Think of this as a different form of negative feedback where the voltage fed back, instead of correcting a discrepancy in voltage, corrects a discrepancy in frequency.



One of the interesting things about a PLL is that, under some conditions, it can recover a signal that is below the noise floor. It can also form the basis of a radio receiver without the use of tuning capacitors and inductors. If a digital frequency divider is incorporated into the circuit, the VCO can be locked to a frequency that is a multiple or submultiple of the input frequency. This is one way of generating a number of precisely controlled frequencies while requiring only one high-precision, single-frequency reference oscillator.

SIGNAL COUPLING IC'S

One way of getting a signal from one point in a circuit to another without an actual connection is to use a transformer. However, ICs give us other options which are less expensive and permit us to couple DC. There are many ways of doing this. The classic method is to pair a LED with a transistor in a single package. That's called an **optocoupler**, which we'll be looking at in chapter 10. It's also called an **opto-isolator** because it's used to transfer a signal while providing electrical isolation.



Recently, newer ways of making an isolating coupler, using capacitance or miniature transformers in conjunction with active circuitry, have been gaining market share. Optocouplers are now just one of many options.

APPLICATION-SPECIFIC IC'S

The short name for an **application specific integrated circuit** is **ASIC**. A circuit to perform almost any function can be assembled out of various IC's and discrete parts, but when enough people want a particular kind of circuit, a manufacturer may make a chip especially designed for that purpose. Most ASICs are either analog or combined analog/digital, because functions that are entirely digital can usually be programmed into standard digital chips. The automotive industry is one place where a lot of ASICs are used.

Sometimes a chip IC and some discrete parts are assembled into a package that, like an IC, is treated as a black box. These are called **hybrid** devices, and they look like IC's on steroids. One of the most common is an audio power amplifier, which contains an IC and a pair of output power transistors. Another example is a type of isolation circuit that is built into a standard IC package, but incorporates a miniature transformer.

There are a great many more types of analog IC's, but it's time to move on. Sometime when you're bored, take a catalog from DigiKey or Mouser and thumb through the IC section to get an idea of the vast array of available devices.

HANDLING AND USE PRECAUTIONS

- Never reverse the power supply connections, sending -DC where +DC should go and vice versa.
- Never feed an input with a voltage higher than the +DC supply pin or lower than the -DC supply pin. Sometimes, people feed a higher signal voltage through a high resistance, using the IC to clip the peaks off. This is poor practice; use diodes.

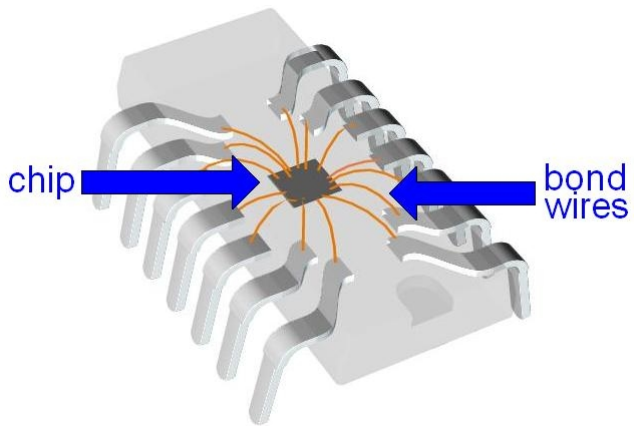
- When soldering an IC onto a PC board, be quick about it. With all those pins, there's a lot of heat, and the circuit can be damaged.
- Some analog ICs use CMOS transistors. As you gain experience with equipment, you'll know which ones those are. Use static-discharge precautions.
- Where ICs are in sockets, never plug or unplug an IC with the power on.
- It's good practice to bypass the DC power connection to the ground rail with an 0.1μfd capacitor very close to the IC.

IC PACKAGES

There are three things an IC package has to do. The first is to seal the circuit so air, moisture and dirt can't get at it. The silicon shapes that make discrete-transistor emitters, bases and collectors are small enough, but with IC's they're microscopic or submicroscopic. It takes only a very tiny amount of contamination to ruin an IC.

The second is to get rid of heat. Most IC's generate only a tiny amount of heat, and it dissipates through the case into the PC board and the air with no problem. Higher amounts of heat can be removed by attaching a heat sink to the package. For IC's that get really hot, a metal plate is molded right into the IC package for efficient heat transfer from chip to external heat sink. It's the same idea as a T0-220 transistor case, and in fact many high-power ICs use a T0-220 case with more than three leads. A computer CPU like a Pentium, when it's working hard, may generate hundreds of watts of heat. To get rid of all that heat, the entire top of the package is made of metal, which is attached to a heat sink, which is attached to a fan.

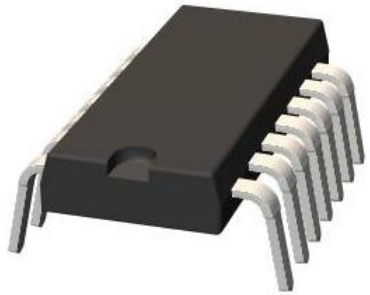
The third thing an IC package does is provide a way to connect the chip to external circuitry. It is impractical to deal directly with the chip in most equipment manufacturing situations. Instead, the IC maker puts the chip in a case with terminals sticking out, and connects those terminals to the chip inside via gold whiskers called **bond wires**. The buyer then takes the IC and attaches the terminals to the external circuit.



One of the main reasons IC's exist is to reduce the physical size of the circuitry. If we trace the IC development over time, we'll see a continuing drive toward ever smaller packages. This only makes sense, because in most cases the size of the actual IC chip is a tiny fraction of the size of the package. The limitation has always been the space needed to attach leads, not the space taken by the chip itself.

The oldest package is a modified T0-5 can with 8 or 10 leads in a circle. For many years this was the only package that was reliably contamination-free. This form has pretty much been replaced by less expensive packages. However, manufacturers don't throw out $100,000 machines every three years. You may very well find yourself repairing equipment that's older than you are. As you gain experience, you will know whether a T0-5 IC can be replaced with a newer plastic part – you'll have to create an adapter so the leads will fit – or whether you need to go online and pay through the nose for an identical IC.

A much cheaper package is a black plastic lump that's molded around the IC. Contamination, which once plastic-encapsulated IC's unreliable, is no longer a problem. For military, high-reliability and high or low temperature applications, a ceramic case is used instead of plastic.



The standard form has long been the **dual inline package** or **DIP**, two rows of leads with the IC package between them. The leads are spaced 0.1-inch apart and the rows are 0.4-inch apart. The most common packages have 8 pins (two rows of 4), 14 pins (two rows of 7), or 16 pins (two rows of 8). Larger packages may have 40 pins or more, and in these cases the rows are farther apart to accommodate larger packages with more circuitry inside. The leads are much sturdier than they would have to be from a current-carrying standpoint, so they can be inserted into PC boards without crunching.

The PC board space can be reduced if we can stand the package on its edge. In this case, we can only put leads along one side. This is called the **single inline package** or **SIP**.

To increase the number of pins without making the package bigger, the package can be made square with pins on all sides. Nowadays, most IC's are soldered directly to the traces on the surface of a PC board, instead of having leads that poke through holes in it. **Surface mount** or **SMT** technology has become dominant.

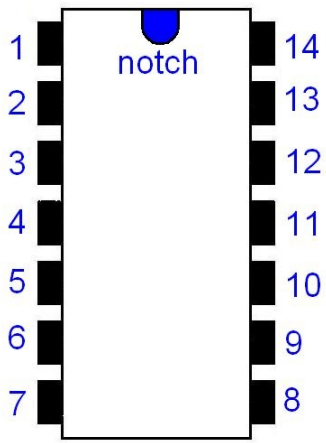


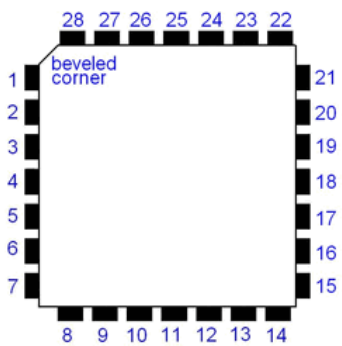
Besides permitting smaller parts, SMT permits parts to be mounted on both sides of a PC board. Without SMT, the smartphone in your pocket would be the size of a shoebox.

SMT component size standards sound like Harry Potter speaking Parseltongue – msop, tsop, ssop, tssop, gfp, plcc/clcc etc. Assembly or service of SMT equipment requires a microscope, special tools, and specialized training.

PIN AND PART NUMBERS

In DIP and most other packages, pins are numbered counter-clockwise from upper left, looking at the top of the IC. A notch or depression tells you which end of the chip to start counting from.





With square packages, look for a depression or a beveled corner.

In diagrams and spec sheets, the positive supply pin is often named V_{CC} or V_{DD} and the negative supply pin V_{EE} or V_{SS}.

IC part designations usually consist of a number that's generic, and two or three letters preceding the number that are associated with a particular manufacturer. For example, the μA741 and the LM741 are the same IC made by different manufacturers. "741" always means a specific kind of opamp, "555" always means a specific kind of timer etc.



Letters after the part number designate the case style, temperature range (military-specified chips have a wider specified temperature range than regular chips) and maybe other details. For example, "CN" means a DIP pin pattern and a plastic case.

An IC also carries a date code, and you have to be alert to tell which is the date code and which is the part number. A designation “1023” isn't a part number, it's an indication that the IC was made during the 23rd week of 2010. The place where this causes the most confusion is in digital ICs found in old machinery. A lot of digital chips have 4-digit part numbers beginning with “74” and “94,” and a lot of those were manufactured in 1974 and 1994. You've been warned. ☺

End of Chapter 5

EM3 Chapter 6
OSCILLATORS

Objectives of this chapter:

- *define "oscillator" and "oscillation"*
- *discuss positive feedback*
- *explain the basic operating principles of tuned and non-tuned oscillators; define "multivibrator"*
- *discuss various forms of LC oscillators: Hartley, Colpitts, Clapp, Armstrong, Meissner, non-tapped tanks*
- *discuss various forms of RC oscillators*
- *explain why AGC is needed in an RC oscillator*
- *discuss crystal oscillators*
- *describe frequency multipliers*
- *describe how various waveforms may be obtained from a multivibrator*
- *discuss single-pulse circuits*
- *discuss noise generators*

Up to this point, we've looked at many ways to amplify, process and transport signals, without bothering to think about where they come from. It's time to learn how some of those signals are generated in the first place.

This chapter deals primarily with the generation of continuous, single-frequency waveforms. There are thousands of places where it's necessary to do that, including places you wouldn't think of. For example, I'm diabetic, and I have to test my blood sugar several times a day. The photograph shows outer and inner views of the tester I use. The tester has a computer chip in it, and the chip has to perform a number of functions in a precisely timed sequence. So, right next to the computer chip, there's a quartz crystal, which is a device commonly used to generate a continuous waveform at a precise frequency.



A circuit that generates a continuous waveform is called an **oscillator**. The up-and-down voltage fluctuations it creates are called **oscillations**.

POSITIVE FEEDBACK

We've seen how, in an amplifier with negative feedback, the signal path is a complete loop from input through the amplifier to output and back to input. We've seen how negative feedback can be used to stabilize a circuit's gain or operating point. With an oscillator, stability is exactly what we don't want. We want the output voltage to cycle constantly, never finding a stable rest point. So, instead of a negative feedback loop, we build a **positive feedback** loop.

You've heard the expression “vicious cycle” used to describe a bad situation or behavior that keeps making itself worse, chronic drinking for example. Positive feedback is like that, but with oscillators it's a good thing. Another expression for negative feedback is **degenerative feedback**, while positive feedback is **regenerative feedback**. It regenerates; the signal feeds on itself. Every change causes a greater change.

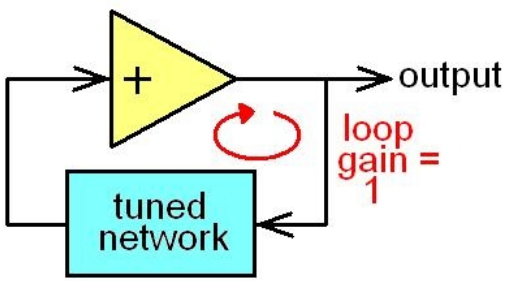
Most oscillators fall into one of three groups: the **tuned** oscillator, the **multivibrator** oscillator, and a partilcular kind of multivibrator called the **relaxation** oscillator.

TUNED OSCILLATORS

In a tuned oscillator, any small noise in the circuit (and there's always atomic noise, at least), is amplified by positive feedback and amplified again, becoming stronger with every trip around the loop, until the whole circuit is chasing its tail, generating a continuous waveform.
For a tuned circuit to oscillate, the conditions are:

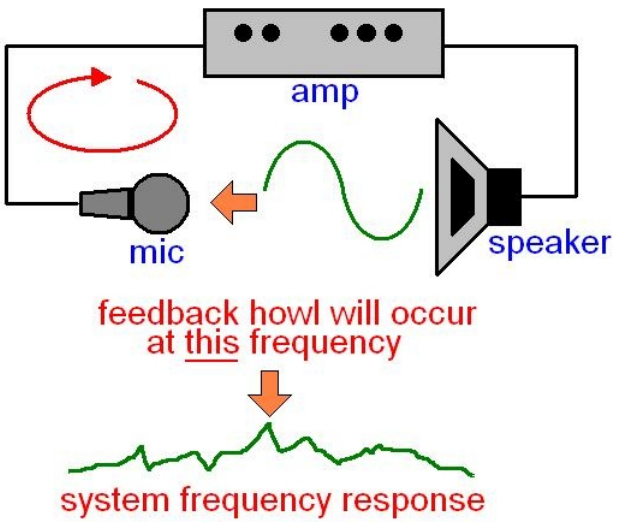
- a positive feedback loop must be present
- at some frequency, gain around the loop, from input to output and through the feedback network back to input, must be one (unity) or greater
- at that frequency, the phase shift around the loop must be zero degrees

These conditions, which are formally known as the **Barkhausen criteria**, insure that during each oscillation enough of the output signal is fed back to the input to “prime the pump” for the next oscillation, so the process keeps going.



In a tuned oscillator, there is *one* frequency where loop gain is unity and phase shift is zero. At all other frequencies, loop gain is less than unity or there is phase shift; thus there is only one possible oscillation frequency. This condition is realized by putting a tuned or frequency-discriminating circuit somewhere in the loop.

If you've ever heard a sound system howl, you've heard positive feedback in a tuned system. Sound comes from the speakers, some of it gets into the microphones and is amplified and comes out the speakers again.



The “tuned” part comes from the fact that there is no such thing as a sound system with completely flat frequency response, especially when you take room acoustics into account. The system will be more sensitive to some frequencies than others, and there is always one frequency where the response is greatest. As the gain is turned up, a point will be reached where, at that frequency, loop gain is just a touch greater than unity. At that point the system goes into self-sustaining oscillation. The frequency of oscillation is called the **resonant frequency**, abbreviated **Fr**.

In a tuned oscillator, the fact that there is only one frequency where loop gain is unity implies that we're dealing with a sine wave. If some part of the circuit is overdriven, then the output at that point will be distorted, which is to say not a sine wave. However, the natural tendency of a tuned oscillator is to produce a sine wave at some frequency, simply because the gain at all other frequencies is less than unity so they are attenuated each time around the loop, leaving us with a waveform that contains energy at only one frequency. The only waveform that meets that condition is the sine wave.

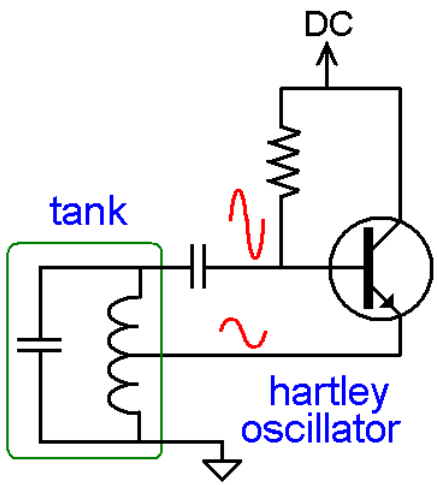
LC OSCILLATORS

The simplest oscillators combine a single transistor and an inductance-capacitance tank. Loop gain is typically much greater than unity at the resonant frequency. The transistor turns on for only a small part of the cycle, but it turns on hard (saturates); in other words, it operates class C. However, the waveform in the tank itself is a relatively clean sine wave, and we can tap off a little of that energy to get a sinewave output. We just have to be careful not to load the tank so heavily that oscillation is snuffed out.

The following circuits use a single NPN transistor, but a PNP transistor or a FET can be used with suitable bias and polarity modifications

THE HARTLEY OSCILLATOR

One way of building an LC oscillator is to create a feedback loop from the emitter back to the base. This is a common-collector circuit, so the output is the same polarity as the input and the feedback is positive.



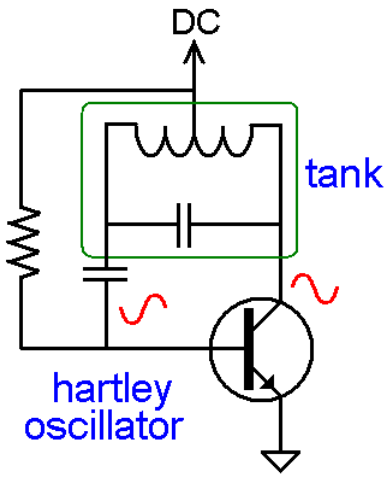
However, the gain of a CC circuit is slightly less than unity, so one of the necessary conditions for oscillation is not met. We can fix this by connecting the emitter to a tap on the inductor. This makes it operate as an autotransformer, stepping up the voltage so the AC voltage at the base is greater than at the emitter. Thus we have a positive feedback loop with a gain greater than one, and the circuit oscillates. An oscillator with a tapped inductor is called a **Hartley** oscillator after its inventor.

What are the extra capacitor and resistor doing there? Well, to get the oscillator going, we have to bias the transistor "on" just a bit. The inductor is connected to ground, and the transistor's emitter is connected to DC ground through the inductor, which means the base must start at +0.7VDC. So we couple the tank to the base through a coupling capacitor, and provide bias current via a resistor.

Once oscillation starts, the transistor will be biased "off" most of the time – it operates class C, remember – and the average voltage on the base will be negative. That's not as strange as it sounds. The base-emitter circuit is a diode, so the waveform at the base can never be more than 0.7V positive at any moment. Every time the sinewave goes positive, current flows through the base-emitter diode and the coupling capacitor charges up, with its negative side toward the base. Within a few cycles, the voltage at the base will be a sine wave that goes 0.7V positive on peaks, briefly turning the transistor on – class C – and spends the rest of the cycle more negative than ground.

The output can be taken from the top end of the coil, from the tap, from a second tap, or from a second winding. The cleanest sinewave comes from the end of the coil, but this is also the highest-impedance point, so loading is an important consideration. Often, a clean sinewave is not important because the signal passes through another tank elsewhere in the circuit which cleans it up.

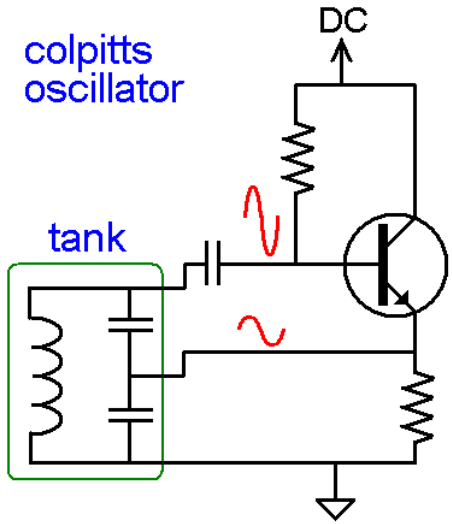
Another way to make a Hartley oscillator is to connect the inductor tap to the DC rail. The collector is then connected to one end and the base is AC coupled to the other. When the collector pulls one end of the inductor in the negative direction, the opposite end of the inductor will go in the positive direction, like a seesaw; thus the inductor performs a phase inversion.



This is a common-emitter circuit, so the transistor performs a second phase inversion and we wind up in-phase, with positive feedback and, since a CE amplifier has voltage gain, an overall loop gain greater than unity.

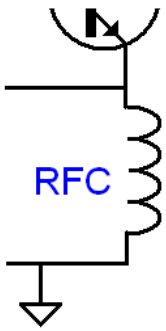
THE COLPITTS OSCILLATOR

Instead of tapping the inductor, we can "tap" the capacitor – that is, we can connect two capacitors in series. The operating principle is the same as with the Hartley oscillator, but it's not quite as obvious. You're familiar with resistive voltage dividers. Series capacitors can be voltage dividers too – AC only, of course. Suppose we replace a tank's capacitor with two capacitors in series. If we don't want the resonant frequency to change, they'll both have to be larger than the capacitor we replaced, so their reactances sum to the same value as before (product over sum formula). Whatever AC voltage exists across the tank circuit will be divided between the two capacitors. So, once again, we have an AC voltage at the base that's greater than the voltage at the emitter, and a loop gain greater than one. An oscillator with series capacitors is called a **Colpitts** oscillator after its inventor.



The tapped-coil Hartley oscillator provides a DC path from emitter to ground through the coil, but the Colpitts oscillator doesn't. So we have to add a current path in the form of an emitter resistor.

Sometimes, an inductor is used in place of the emitter resistor. This inductor has nothing to do with tuning. It has low DC resistance but high reactance at the oscillator frequency. Its function is to provide a DC path to ground

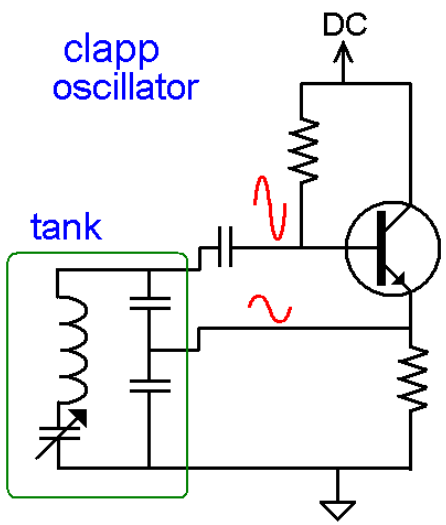


without loading the AC oscillations as a resistor would. When an inductor is used to block AC, it's called a **choke**. If the frequency of the AC it's blocking is high enough to be in the radio band, the choke is called a **radio frequency choke** or **RFC**.

Memory aid: the Hartley oscillator uses a tapped inductor. Inductance is measured in **H**enrys, abbreviated **H**, as in **H**artley. The Colpitts oscillator uses a tapped capacitance, abbreviated **C** as in **C**olpitts.

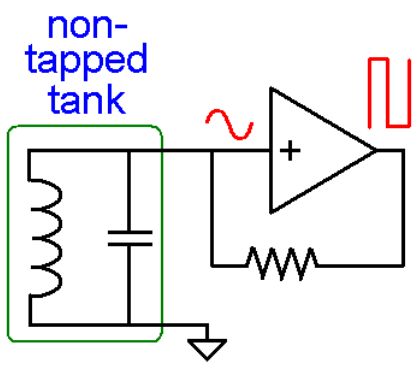
THE CLAPP OSCILLATOR

If we want to vary a Colpitt oscillator's frequency, we can do it by varying one of the capacitors in the voltage divider. However, this also changes the voltage divider ratio. The **Clapp** oscillator is a variation on the Colpitts that solves the problem by adding a variable capacitor in series with the inductor. This permits us to change the tuning without changing the capacitor divider ratio.



NON-TAPPED LC OSCILLATORS

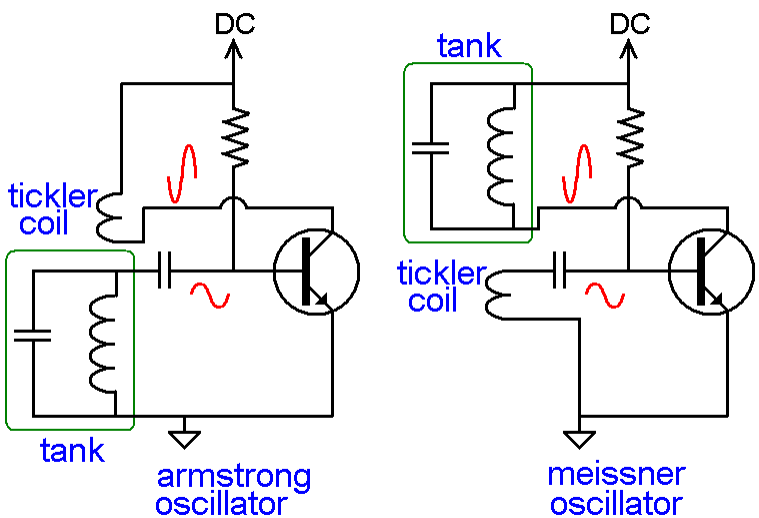
If you have a non-inverting amplifier with a gain higher than unity – two transistors or an opamp, say – neither of the tank components needs to be tapped. Positive feedback creates a loop gain greater than unity, while the tank shunts feedback and shifts phase at every frequency except its resonant frequency, leaving only one frequency where the Barkhausen criteria are met. The waveform at the amplifier output will probably be a square wave, assuming the gain is fairly high, but the waveform across the tank itself will be a sine wave.



TRANSFORMER OSCILLATORS

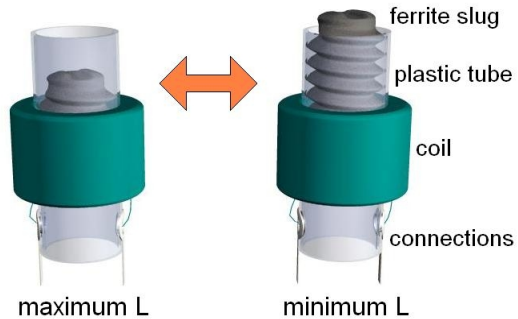
Feedback can be supplied by a separate winding on the coil. It's cheap and easy, just a few extra turns of wire. This is known as a **tickler coil**.

In the illustration below, the left-hand oscillator is called an **Armstrong oscillator** after Edwin Armstrong, better known as the inventor of FM radio. If the tank and tickler are swapped, it becomes a **Meissner oscillator**, though it's probably more often thought of as an Armstrong oscillator. In both cases the magnetic coupling is loose; that is, the mutual inductance is small. If gain or coupling is reduced so the circuit does not quite oscillate, it becomes a very sharply tuned **filter**, which we'll study in chapter 8. The first use of the Armstrong circuit was as a filter in another of his inventions, the regenerative receiver.



TUNING AN LC OSCILLATOR

The frequency of an LC oscillator is changed by adjusting either the tank's capacitor or its inductor. Variable RF inductors are typically built with ferrite cores that can be screwed in or out to change their value. Variable capacitors have a shaft for a knob if they're going to be frequently changed; otherwise they're adjusted with a screwdriver. When an oscillator is going to be adjusted just once, during manufacture, it's customary to **dope** the variable capacitor or inductor after it has been adjusted, which means putting a dab of glue on it so it can't go out of adjustment.



CRYSTAL OSCILLATORS

In a crystal oscillator, resonance is not achieved by an LC tank, but by the physical vibrations of a piece of crystal. Certain materials, notably quartz, have a **piezoelectric** effect. If they are flexed, they put out a voltage. If a voltage is applied across them, they flex. Thus a quartz crystal lets us change an electrical signal to a mechanical vibration and back to an electrical signal. "Crystal" is often abbreviated **XTAL**.

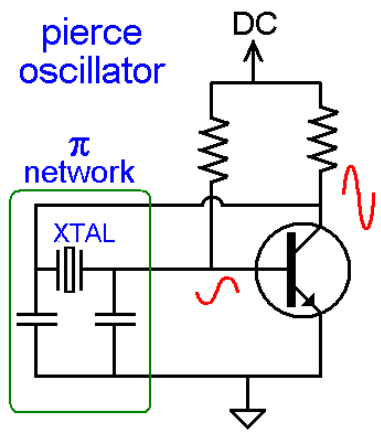
A crystal has a natural frequency of vibration, like a chime. If we attach a crystal to an amplifier which both feeds energy to it and draws energy from it – if we make it part of a positive feedback circuit – we have the mechanical equivalent of a tuned LC circuit.

The crystal oscillator's great virtue is stability. Crystal oscillators have always been used where it is necessary to obtain a very precise frequency. Only two things throw them off. One is a change in circuit capacitance. This isn't a big problem; in most circuits, capacitance is fixed.

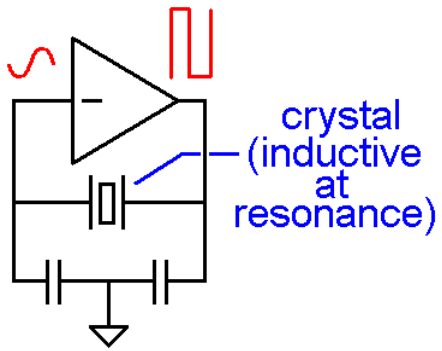
(Sometimes a variable capacitor is used to adjust the frequency. This is called **rubbering** the crystal. Since the crystal is the dominant factor in setting the frequency, the available range of adjustment is very small.)

The other thing that can throw off a crystal oscillator is temperature change. In most applications, the error is so small that it doesn't matter. If necessary, the crystal can be placed in a miniature, thermostatically controlled oven to eliminate even this small variation. Accuracy of ±1Hz in 10^8 (0.000001%) is typical for a temperature-controlled crystal oscillator.

The most common crystal oscillator arrangement is known as a **Pierce** oscillator. At its resonant frequency, the crystal has the electrical characteristics of an inductor, which makes this a variation on the Colpitts oscillator. The crystal and the capacitors on either side form a pi (π) network with a resonant frequency almost completely governed by the crystal's resonance. In a pi network, there's a signal inversion between ends. The CE transistor amplifier provides another inversion, and we wind up with the zero degree overall phase shift we need for oscillation.



A problem with the transistor oscillator is that the base loads the resonant circuit, which can lead to quirks including unreliable oscillation. This can be avoided by using a FET, which loads the signal input very little. Another option is to use an integrated amplifier, which typically has a very high input impedance.



An oscillator that is closely related to the crystal oscillator uses a **ceramic resonator**, a piece of ceramic material that serves as the resonating "chime," bonded to a crystal that provides the necessary piezoelectric effect.

RC OSCILLATORS

Below a few hundred kHz, the size and cost of large, magnetically shielded inductor is prohibitive. Also, in many applications the oscillator frequency must be variable by as much as a ten-to-one ratio. In an LC oscillator, frequency change is proportional to the square root of inductance change or capacitance change. That means, if we need a 10:1 frequency ratio, either the inductor or the capacitor has to be precisely, reliably variable over a 100:1 ratio. That's difficult to achieve.

For these reasons, various kinds of oscillators have been developed using RC networks. RC oscillators are much less expensive at low frequencies, and the relationship between frequency change and the necessary change in either R or C is a linear relationship, not a square one – a 10:1 frequency change requires only a 10:1 change in R or C. This is due to the fact that, in an LC tank, the reactance of both components is frequency-dependent, but in an RC circuit, only the capacitive reactance changes with frequency; R is just R at all frequencies.

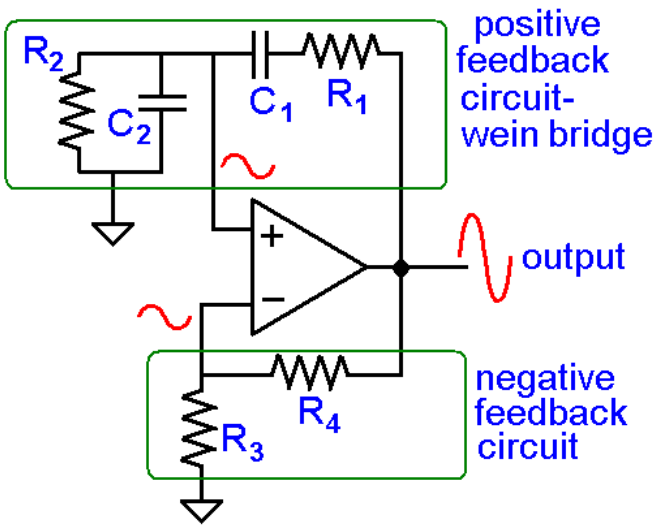
The big problem with RC oscillators is gain stability. In an LC tank, energy is shuttled back and forth between the inductor and capacitor, and a sine wave automatically results. The active device (transistor or IC) can work in class C, as a switch. But in an RC oscillator, the resonant circuit is only very broadly tuned. In some cases, there is no tuned circuit at all: the resonant frequency is determined by phase shift. The active device must work in a linear fashion, which means class A, and gain around the loop must be unity – precisely, absolutely, perfectly.

It is impossible to set up loop gain that precisely. Instead, an RC oscillator includes an **automatic gain control** or **AGC** circuit that constantly monitors the output voltage.

When the AGC senses that the output voltage is greater or smaller than it should be, it makes a slight change in loop gain until the output returns to its proper voltage. Think of the cruise control in a car.

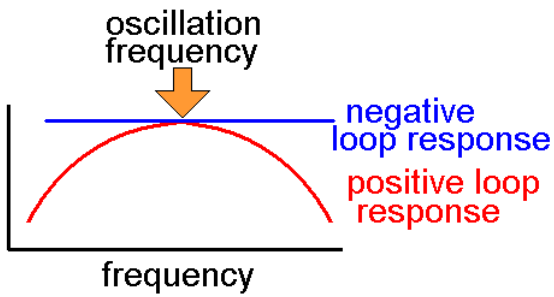
THE WEIN BRIDGE OSCILLATOR

The Wein-Bridge oscillator uses two identical resistors and two identical capacitors in a positive feedback circuit. As shown in the diagram, R_1 and C_1 are in series, and R_2 and C_2 are in parallel.



At high frequencies, there is very little positive feedback because C_2 shunts it out. At low frequencies, there is very little feedback because C_1 blocks it. But at some frequency in between, feedback reaches a maximum of about 30%, and at this frequency the phase shifts caused by the two capacitors also cancel out.

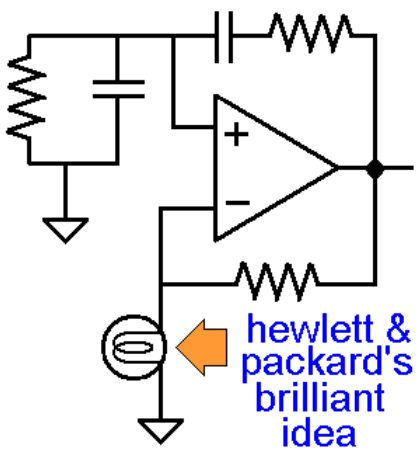
Op-amps make excellent RC oscillators. The fact that we have inverting and noninverting inputs makes it easy to have both positive and negative feedback loops. The positive loop is a Wein bridge, which feeds about 30% of the output voltage back to the "+" input with zero phase shift at one frequency, less voltage and with non-zero shift at all other frequencies. The negative feedback loop is a resistive voltage divider that feeds about 30% of the output back to the "-" input at all frequencies. The two feedback loops balance at one frequency only, resulting in unity gain at that frequency only. At all other frequencies, negative feedback is greater than positive feedback and loop gain is less than unity.



So much for theory; now for the inconvenient reality. Suppose we're building a 1kHz oscillator, and we get it oscillating at 1V. If the loop gain is higher than unity by as little as 0.001%, in about three seconds the signal will grow to the point of clipping. If the gain is less than unity by 0.001%, in about three seconds the signal will die away.

What we need is a way to vary one negative feedback resistor or the other to compensate for output voltage deviations. If we can make a circuit that does this automatically, our stability problem is solved.

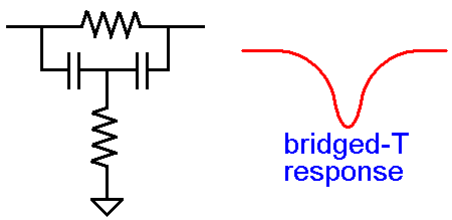
The earliest such device was, believe it or not, an incandescent light bulb. The resistance of a filament is much higher when it's hot than when it's cold. If we put a light bulb in the feedback loop in place of R_3, and work out all the circuit characteristics properly, then any increase in output voltage will put more voltage across the bulb, which will heat the filament, which will increase the resistance, which will decrease the circuit gain. Thus any variation in sinewave voltage is automatically corrected.



Sometimes the most brilliant ideas are the simplest. This particular one was figured out around 1938, only a few years after negative feedback was formally defined, by a guy named Hewlett. He teamed up with a guy named Packard and they formed a company you might recognize.

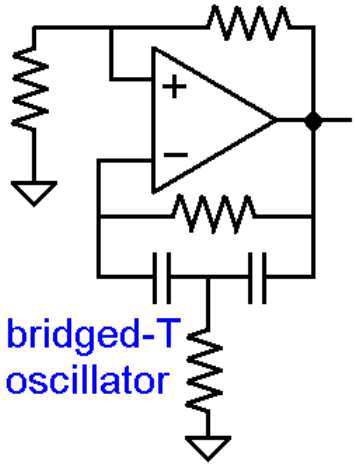
Nowadays, we don't use light bulbs. FETs act like resistors as long as the source-drain AC voltage isn't too great, and they can be varied by changing the gate voltage. The circuit's a little more complicated but not much.

THE BRIDGED-T OSCILLATOR



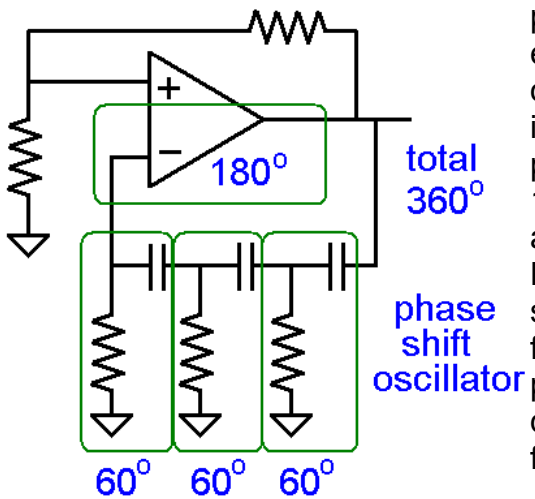
In contrast to the Wein bridge, which has a broad peak in its response, the bridged-T exhibits a notch in its frequency response. There are two current paths through the circuit, and at one particular frequency the phases are opposite, causing partial cancellation. This is the reverse of the situation with a Wein bridge, so we put the bridged-T in the negative feedback loop. At one frequency, negative feedback is at a minimum, and if we adjust the positive feedback loop correctly, we get oscillation.

The circuit at right is a bridged-T oscillator, with negative feedback that varies with frequency and positive feedback that does not. In an actual oscillator, the shunt resistor in the positive feedback circuit would be replaced with a FET, whose value would be varied to provide the AGC function.



THE PHASE SHIFT OSCILLATOR

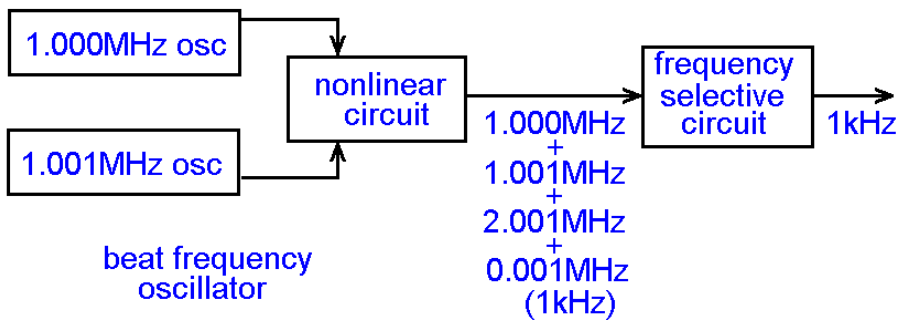
The phase-shift oscillator does not use a tuned circuit. Instead, it uses a feedback network typically incorporating three RC circuits which shift the phase 60 degrees each, for a total shift of 180 degrees. The inverting amplifier provides the other 180 degrees to make an in-phase loop. An RC circuit's phase shift varies with frequency, so the total phase shift is zero degrees at only one frequency. Again, in a real-world oscillator the shunt resistor in the positive feedback loop would be replaced by a FET to provide AGC. Though the phase shift network is not tuned, the complete circuit is resonant, so this counts as a tuned oscillator.



THE BEAT FREQUENCY OSCILLATOR

When two frequencies are fed through a nonlinear circuit, two additional frequencies are created. This is called **intermodulation**. These new frequencies are the sum and the difference of the original frequencies.

For example, if we intermodulate 1MHz and 1.001MHz, we will create signals at the sum, 2.001MHz, and the difference, 1kHz. Intermodulation is a basic part of most radios, so we'll be talking a lot about it in EM6.



One of the interesting things about intermodulation is that a small change in one of the original frequencies results in a large change in the difference frequency, proportionally speaking. For example, if we change the 1.001MHz oscillation to 1.002MHz, an 0.1% change, the 1kHz resultant will change to 2kHz, a 100% change.

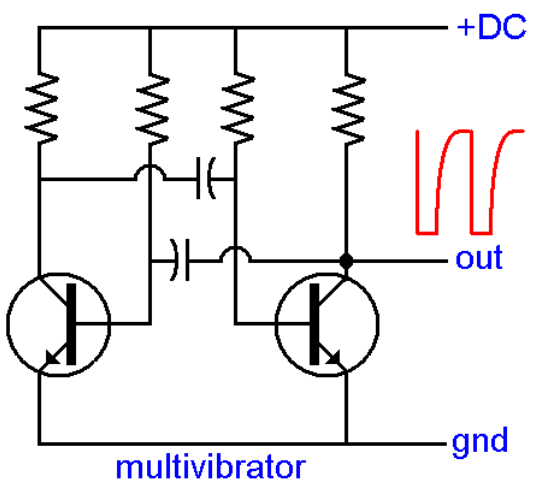
We can use this technique to build a variable-frequency oscillator that covers an extremely wide frequency range in a single sweep. Suppose we want a generator that sweeps the audio band from 20Hz to 20kHz. This is a 1,000:1 range. It is impractical if not impossible to build an RC oscillator whose frequency varies by 1,000:1. Usually, the only way to do that is to break that range into three sub-ranges with 10:1 ratios. However, if we intermodulate a sine wave at 100,000Hz with another that varies between 100,020Hz and 120,000Hz, we only need a capacitor that varies by 20%, which is a walk in the park. Then all we need is a filter that removes the higher frequencies and leaves us with the lower-frequency intermodulation result. The two high-frequency oscillators "beat" against each other and the high and low frequency resultants are called **beat frequencies**. The circuit in its entirety is called a **beat-frequency oscillator** or **BFO**.

All the oscillators we've looked at use resonance to generate sine waves. The other kind of oscillator has no tuned circuits; frequency is determined by the charging and discharging time constants of RC circuits. This kind of oscillator is called a **Multivibrator**. In most cases the output resembles a square wave.

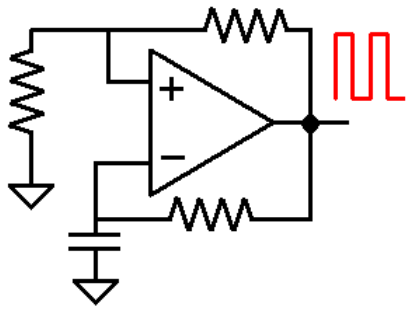
MULTIVIBRATORS

A multivibrator is often referred to as an **astable** ("A-stable," long A) circuit. The prefix "a" means "without"; e.g. a person with no morals is amoral. An astable circuit has no stable state. It constantly switches between two states, never settling on either one. A buzzer is a mechanical version of an astable multivibrator.

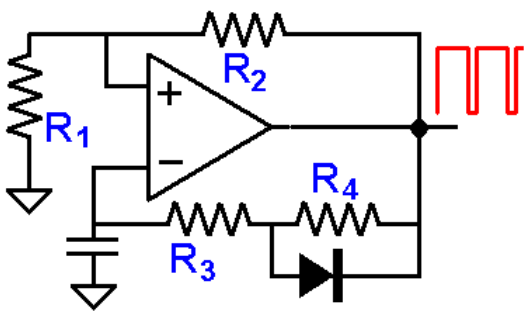
There are any number of ways of making a multivibrator. A simple one can be made from two transistors. I wrote a detailed explanation of its operation and then deleted it, because it ran three full pages. I challenge you to figure it out yourself.



Hint: note that the collector of each transistor is capacitively coupled to the base of the other. Both are CE circuits, meaning we have two polarity reversals, meaning the feedback is positive. Another hint: the capacitors alternately drive the two bases negative with respect to ground. Third hint: study the waveform carefully.



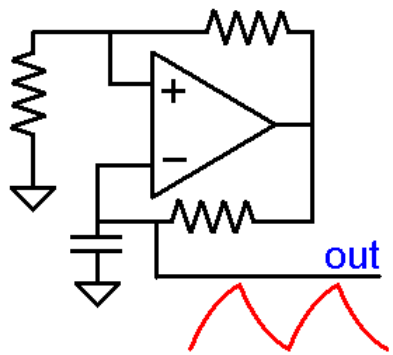
Opamps can make multivibrators, using both positive and negative feedback loops. In this circuit, note that the positive feedback circuit is a voltage divider, while the negative feedback circuit has an RC time constant. The positive feedback circuit drives the output one direction; then, when the RC circuit catches up, reversing the polarity of the voltage across the inputs, the positive feedback circuit drives the output the other direction.



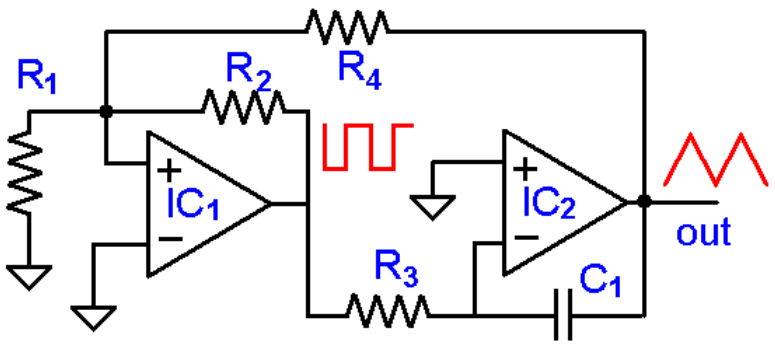
If a diode is added to one of the feedback circuits, we can make a multivibrator with a duty cycle that's far from 50%. The RC time constant involves only R_3 in one direction, and in the other, R_3 and R_4 in series. Thus the positive-going and negative-going time constants are not the same.

THE TRIANGLE WAVE GENERATOR

The circuits above have square or rectangular wave outputs. The multivibrator is also the basis of the triangle wave generator. The capacitor charges up and down during each cycle, and we can take the output from this point if we want a triangle wave.



The simple circuit at right works for many applications; I often use it myself. But it has limitations. One, if we load the circuit we'll upset it, so we have to run the signal through a voltage follower. Two, the waveform is not really a triangle wave, because we're dealing with an RC charge/discharge. If we want linear up-and-down ramps, we'll have to replace the feedback resistor with a constant current source.

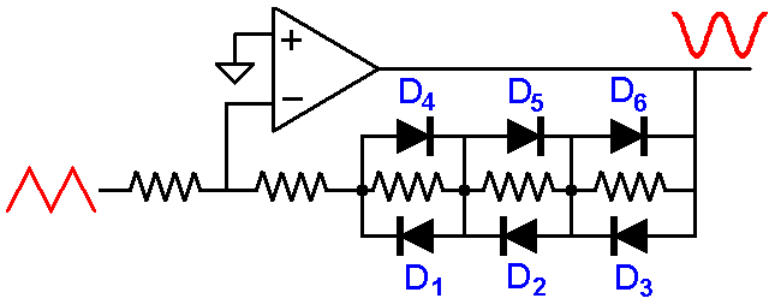


The circuit above solves both problems. IC_1 outputs a square wave, which supplies unvarying current, alternately positive and negative, to the virtual-ground input of IC_2 via R_3. This constant current causes C_1 to charge up and down at a constant rate. The positive feedback loop of R_1, R_2 and R_4 insures that the circuit never finds a stable resting point and continues to oscillate. Finally, the signal is taken from the output of IC_2, a low-impedance point.

If we use the diode-and-resistor trick around R_3, we can move the duty cycle away from 50% to make a sawtooth generator.

THE MULTIVIBRATOR SINE WAVE GENERATOR

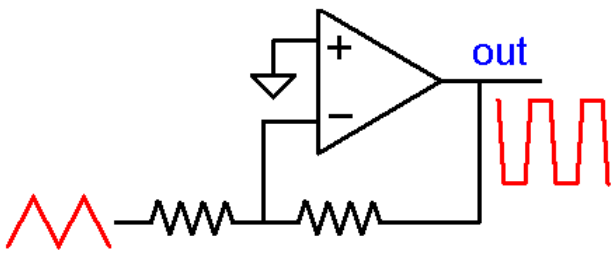
In the circuit above right, a triangle wave is fed to a second circuit with a nonlinear feedback loop. As the output voltage rises on the positive alternations, diodes D1, D2 and D3 begin to conduct, one after another, reducing the feedback resistance. On the negative alternations, D4, D5 and D6 do the same thing. This squashes the peaks of the triangle wave, resulting in something approaching a sine wave. This is often called a **piecewise** sine wave because it's made up of linear segments rather than a true, smooth sine curve.



A popular and useful test instrument is the **function generator**, which allows the user to select square, triangle or piecewise sine wave output. Because the waveforms are not generated by tuned circuits, it's easy to obtain a stable signal voltage over a wide range of frequencies; no AGC required. The main drawback is that the sine wave is not free of harmonics. You can use it to measure the frequency response of a system, but you can't use it to measure distortion because its signal is a little distorted to begin with. A decent piecewise-generated sinewave has around 1% distortion.

TRAPEZOID WAVE GENERATORS

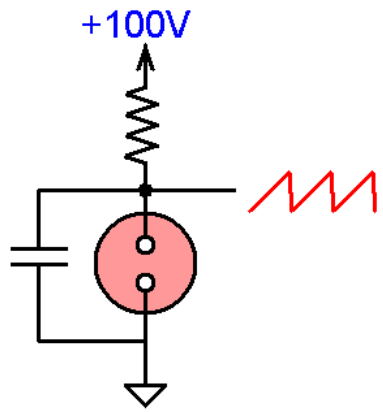
A trapezoid wave is a square or rectangular wave with stretched-out rise and fall times. It is used as an approximation of a sine wave, or where it is necessary to reduce the RF radiation associated with fast rise and fall times. One way of generating it is to feed a triangle wave into an amplifier and set the gain high enough that the amplifier clips.



RELAXATION OSCILLATORS

Relaxation oscillators are like multivibrators in that they use an RC time constant to determine frequency. In a relaxation oscillator, each cycle consists of a short burst of current followed by a much longer period where the circuit coasts or "relaxes." Usually, we're talking about a capacitor that is charged up slowly and discharged quickly.

The Neon Lamp Oscillator



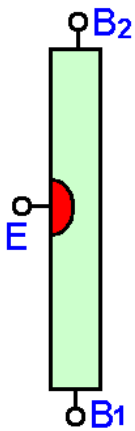
The simplest and probably oldest relaxation oscillator uses a small neon lamp. Neon lamps have an interesting current characteristic: they require a certain voltage to "fire," but once they're glowing, they like to operate at a lower voltage. Typical voltages are 70V to fire and 60V to operate. When a neon lamp is used as a pilot light in a 120VAC circuit, a dropping resistor limits current.

If a capacitor is added across the neon lamp and a high enough DC voltage is applied, the circuit will begin to oscillate. The cycle works like this:

1. The capacitor charges up to the firing voltage and the lamp fires.
2. The lamp discharges the capacitor to the operating voltage.
3. The resistor, which has a very high value, cannot supply enough current to keep the lamp lit, so it goes out.
4. Repeat.

This circuit was very popular at one time, as it is stupid-simple and dirt cheap. The main drawbacks are that the frequency isn't very stable, and nowadays most equipment doesn't use tubes, so there isn't a high voltage supply available.

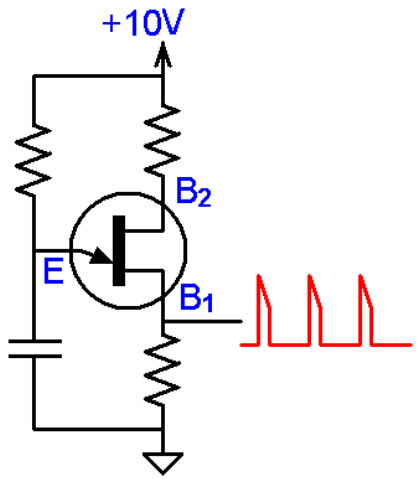
The Unijunction Oscillator



A relaxation oscillator that operates on lower voltages can be built with a **unijunction transistor** or **UJT**. It's called a unijunction because it has just one PN junction. As noted in chapter 2, it consists of a tiny slab of lightly-doped N-type silicon with a base connection at either end and a PN junction somewhere along its length, which consists of a heavily P-doped emitter. If we start with a low emitter voltage and let it rise, when it gets to 0.7V more positive than that point in the base, the UJT will fire.

A UJT oscillator consists of resistors that connect the bases to ground and +DC respectively, and an RC charging circuit. When the capacitor charges to the required voltage, the UJT becomes a near-short circuit, putting out a pulse and discharging the capacitor.

The charging resistor has a high enough resistance that, once the capacitor is sufficiently discharged, it cannot supply enough current to keep the UJT in conduction, so the UJT resets and the cycle starts again. Aside from the slightly more complicated circuit, this works exactly like the neon oscillator.



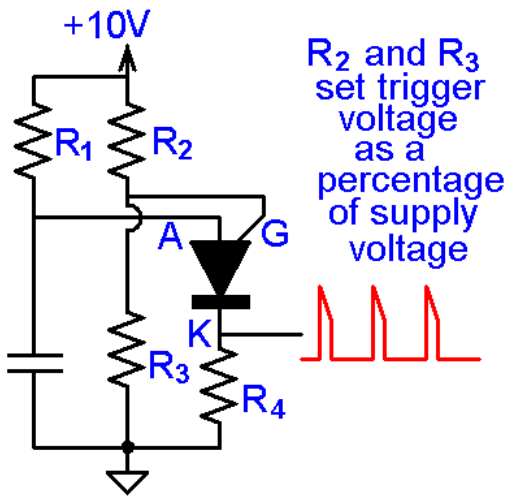
UJTs have not been manufactured in decades. If you're working on an old machine with a burned out UJT, you probably won't be able to find a replacement.

However, there's a closely related device called a **programmable unijunction transistor** or **PUT** that is still made and still available from places like Mouser and Digikey. The only real way to replace a UJT these days is to modify the circuit so as to use a PUT.

The Programmable Unijunction Oscillator

The phrase “programmable unijunction” is a contradiction: a PUT has three junctions, so it's not a UJT. But it works like a UJT so that's what people call it.

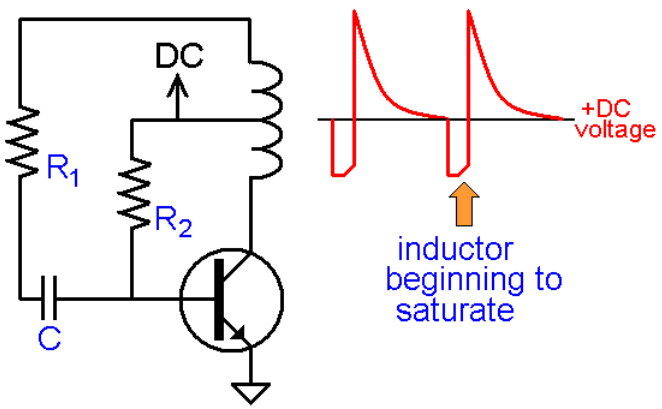
The difference is, in a UJT the trigger voltage is determined by its construction, but with a PUT it is set by a resistive voltage divider. That means you can put a PUT where a UJT used to be, if you're willing to add two resistors. As the symbol in the diagram shows, a PUT is electrically much more closely related to a thyristor than a UJT. However, once the trigger voltage is set by the voltage divider, it behaves almost exactly like a UJT.



THE BLOCKING OSCILLATOR

A blocking oscillator can be made with a FET, but its most natural configuration uses a BJT. Operation is simple, but describing the operation is not. Here's my best shot.

At power-up, R_2 turns the transistor on slightly. The low end of the inductor is pulled down, causing the high end to rise. This rise is coupled to the base of the transistor, turning it all the way on. You'll recognize this as positive feedback. C charges up, with positive voltage to the left.

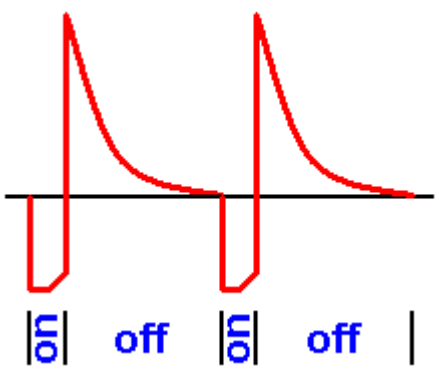


With a voltage across the inductor, current begins to rise, along with its magnetic field. Pretty soon, it saturates, and the high current pulls the transistor out of saturation. Voltage across the inductor starts to fall; the collector end goes in the positive direction and the base end goes in the negative direction. Since there's a charge across the capacitor, its right side, which iwas at 0.7V, begins to fall, turning the transistor off.

At this point, flyback action reverses the voltage across the inductor. The end connected to the collector goes higher than the DC supply voltage, while the base is driven more negative than ground.

After a time, current through R2 charges C up to the point where the transistor is again biased on, and the cycle repeats.

Note that, despite the fact that the circuit has an inductor and capacitor, there is no tuned circuit. The operation is controlled by the inductor's saturation time and the RC time constant, so it's a relaxation-type multivibrator. It's a crude device whose frequency can only approximately be predicted.



The blocking oscillator has that name because the transistor spends the majority of each cycle "off," with the transistor not conducting ("blocked," in archaic vacuum tube terminology). You could argue that Hartley, Colpitts and all other class-C oscillators are, by this definition, blocking oscillators. But those use tuned circuits; when people say "blocking oscillator" they're talking about a kind of multivibrator.

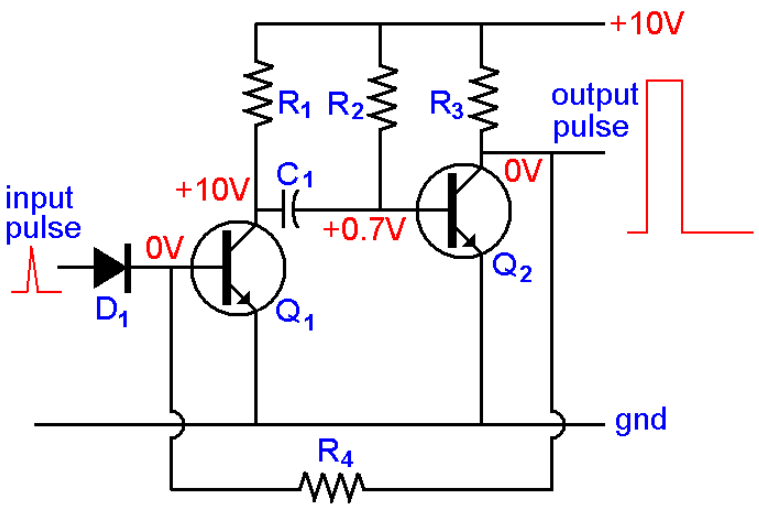
A blocking oscillator is often used where DC must be turned into AC and the exact frequency doesn't matter. Due to its simplicity and high efficiency, it's popular where a high DC voltage (after rectification) must be obtained from a lower DC voltage. Wristwatch displays and xenon flash units typically use blocking oscillators.

PULSE GENERATORS

If we change an astable multivibrator into a circuit that has one stable state, it becomes a **monostable** multivibrator. It normally rests in its stable state. In response to an input pulse, it begins to oscillate, flipping to the other state and back again. However, once it returns to the initial state, oscillation stops.

A monostable multivibrator is not, strictly speaking, an oscillator. I'm discussing it here because the circuit is so closely related to a multivibrator oscillator. The usual purpose for a monostable multivibrator is to create a single pulse of specific duration in response to an input signal. For this reason, it's often called a **one-shot** circuit.

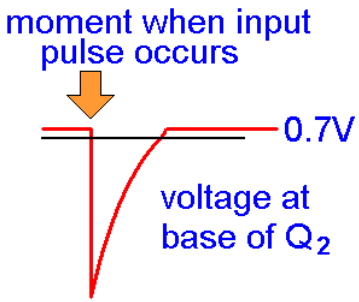
A simple one-shot is shown below. Initial voltages are as shown. Q_2 is biased fully on by R_2. Since its collector is near 0V, there is no current through R_4 so Q_1 is off. Note that there is 9.3V across C_1.



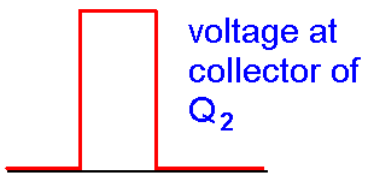
When an input pulse occurs, Q_1 conducts and its collector goes to 0V. That drives the right side of C_1 negative by 9.3V, turning off Q_1 and allowing its collector to go high. Once in this state, Q_1 is kept "on" by R_4, even in the absence of further encouragement from the input.

R_2 pulls up on the right side of C_1, discharging the voltage across it. When it reaches 0.7V, Q_2 turns on, Q_1 turns off, and the circuit is reset.

Note that it makes no difference how long the input pulse is. If it is longer than the output pulse, Q_2 remains saturated, but the base of Q_1 still goes 0.7V positive and Q_1's collector goes to zero. The output pulse duration is set by R_2 and C_1 and no other factor plays a part.



Got all that? If not, trace the circuit out for yourself until you can visualize its operation and "see" the voltages going up and down.



I'll mention just one practical application. It's often nice to know when a short-duration signal peak comes along, but a meter won't show it – too sluggish. The usual way of showing a peak is to have it light an LED. But some peaks are even too fast for an LED to indicate. A 1-shot can be used to stretch out a peak long enough for you to see it.

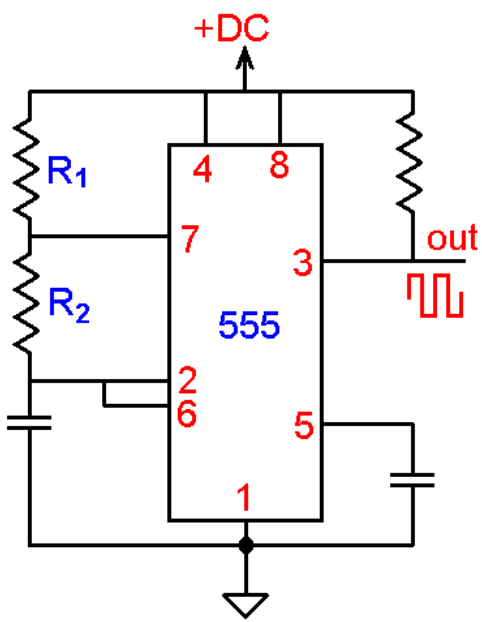
BISTABLE CIRCUITS

Just to round out our discussion of multivibrators: we've discussed the astable multivibrator, which is not stable in either state, continually oscillating between them, and the monostable multivibrator or one-shot, which is stable in one state; if triggered into the second state, it will return to the first state and stay there as soon as its RC time constant times out. There's also such a thing as a **bistable** (BUY-stable) multivibrator, which is stable in either state, like a light switch. In EM4, we'll study **latches**, which are bistable multivibrators by another name.

THE 555

The 555 isn't the only timer/oscillator chip, but it's one that's so widely used I couldn't omit it from this chapter. I'm tempted to give it several pages, but instead I'll just hit the highlights and refer you to the internet. Millions of 555s are made each year, and a tech is guaranteed to run across equipment that uses them.

The circuit at right is a 555 multivibrator. The RC time constant is determined by R_1, R_2 and C_1. IC pin 6 is a trigger that fires when C_1 charges up to about $^2/_3$ times the +DC voltage, pulling pin 7 to ground. This discharges C_1 through R_2. Pin 2 is a threshold detector that fires when C_1 discharges to about $^1/_3$ the +DC voltage, releasing pin 7 and allowing C_1 to charge up again. Pin 5 is an internal voltage reference, bypassed by C_2. The output is taken from pin 3. This is an excellent circuit to practice visualizing voltages going up and down.



Duty cycle depends on the ratio of R_1 to R_2. If R_2 is much smaller, the output will be very short, negative-going pulses.

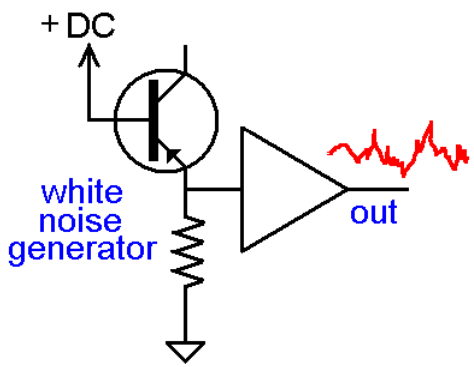
Other things a 555 can do:

- Frequency modulation
- One-shot
- Pulse-width modulation (class D)
- Switch debouncing (we'll run into this in EM4)
- ...and so on

The 555 is so versatile and can do so many things besides oscillating that it would be worth downloading some application notes and studying it closely. Though it just passed its 40th birthday, it's still the standard timing chip.

NOISE GENERATORS

A circuit that generates white noise is not, strictly speaking, an oscillator, but it is a signal generator. There are two common ways of generating white noise. One is to obtain a noisy device, extract a signal from it, and amplify it. A reverse-biased PNP base-emitter junction is often used for this purpose. The result is truly random noise. The other it to generate "imitation" white noise, called **pseudo-random** noise, involving digital techniques.



A natural noise source creates **white noise**, which contains an equal amount of energy across every equal span of frequencies, e.g. the same amount of noise between 5kHz and 15kHz as between 80kHz and 90kHz. Sometimes we need a different energy distribution. That is obtained by taking white noise and sending it through a frequency-selective circuit called a **filter,** which we'll study in chapter 8.

ALMOST-OSCILLATORS

Speaking of filters – if we adjust a tuned oscillator so that its gain is a tiny bit less than unity, it will stop oscillating, but it will be extremely sensitive to an input signal at its almost-oscillation frequency. Major Armstrong used one of his oscillators, adjusted so it was just shy of oscillation, as a radio receiver. This is called a **regenerative receiver** and in its day, the early 1920s, it was much better at picking one station out and rejecting others than its simple design would suggest. We'll be studying regenerative receivers in EM6.

In the meantime, now that you've learned how to generate sine waves with tuned circuits, and square waves and pulses with multivibrators and relaxation oscillators, we're going to spend the next chapter modifying these waveforms.

End of Chapter 6

EM3 Chapter 7
WAVE SHAPING CIRCUITS

Objectives of this chapter:

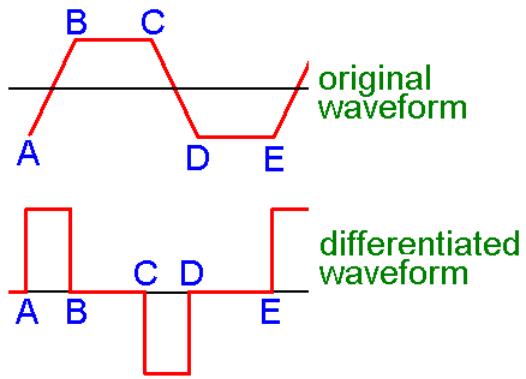
- *explain differentiation and integration*
- *explain practical limits of circuits that differentiate and integrate*
- *discuss RC differentiation and integration circuits*
- *discuss opamp differentiation and integration circuits*
- *describe various waveform clipping, clamping and rectification circuits*
- *give examples of conversion from one waveshape to another*

In electronic equipment, certain functions involve waveforms derived from other waveforms. In this chapter, we're going to study circuits that reshape and modify the standard waveforms we've studied up to this point.

DIFFERENTIATION

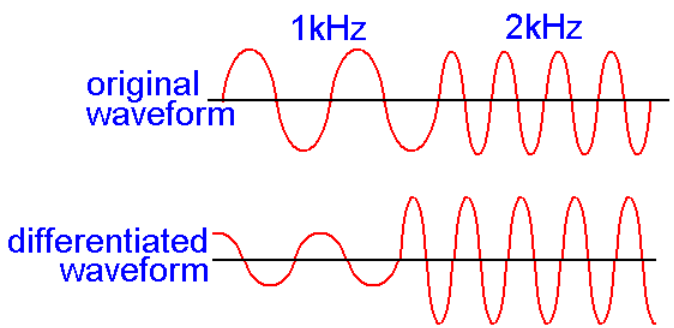
A differentiator takes a waveform and extracts information about how the waveform is changing over time. Its output at a given instant represents not the voltage of a signal but the *rate of change* of that voltage.

For example, consider a trapezoid wave. The voltage ramps up at a constant rate between A and B, does not change between B and C, ramps down at a constant rate between C and D, and does not change between D and E.



If we differentiate this voltage, we get the lower waveform in the illustration. Between A and B, the positive voltage corresponds to the positive voltage change in the original waveform. Since the rate of change is constant, the differentiated quantity between A and B is constant. Between B and C, since the original waveform does not change, the differentiated voltage is zero. Between C and D, since the original waveform is changing in the negative direction, the differentiated waveform is negative. And, between D and E, since the original waveform is not changing, the differentiated voltage is zero. Thus the differentiated waveform, at every point, represents the *rate of change* in the original waveform.

When a sine wave is differentiated, the result is also a sinewave, but its amplitude is proportional to frequency; that is, the higher the frequency, the greater the output. Suppose we start with a 1kHz sinewave and then, without changing the amplitude, we raise the frequency to 2kHz. The voltage moves through the same range up and down as before, but it has to do it more quickly to cover that range in less time. (You remember that, as a waveform's frequency goes up, its period goes down.) That means that the change in voltage during any given period is greater, so the differentiated waveform is larger. It also leads the original signal waveform by 90°.



Mathematically, the instantaneous output of a differentiator is equal to change in voltage per change in time:

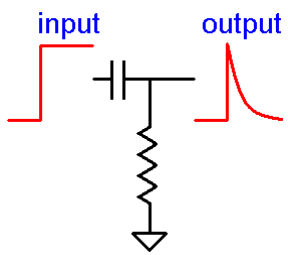
$$V_{OUT} = \Delta V_{IN} / \Delta T$$

Depending on the circuit, the differentiator may also have overall voltage gain or loss, which simply means the result from the calculation must be adjusted up or down.

The output signal of a differentiator is referred to as the **derivative** of the input signal.

The Simplest Differentiator

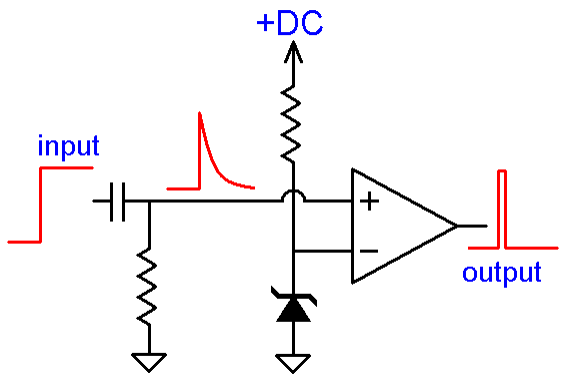
It's customary to be pretty sloppy about the definition of "differentiation." In many cases we have no choice, because theoretically perfect differentiation is often impossible. In practice, any circuit that performs according to the general idea of differentiation qualifies as a differentiator.



The simplest "almost-differentiator" is a capacitor and a resistor. We already know this circuit as an AC coupling or DC blocking circuit. The capacitor passes voltage changes but blocks DC. If we apply a voltage that changes once, instantaneously – called a **step** change – the output is a spike that returns to approximately zero in five RC time constants. The capacitor is made small enough that, in response to a step input change, the output voltage very quickly falls to zero.

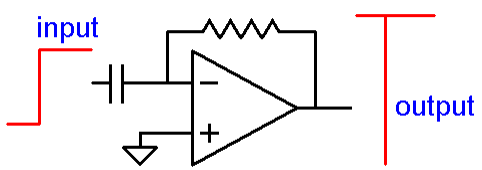
Even though this doesn't meet the theoretical definition of a differentiator, it's often good enough. Usually, what we're trying to do with a differentiator is to identify the moment when a voltage suddenly changes. The RC circuit does that as well as a more complicated circuit would.

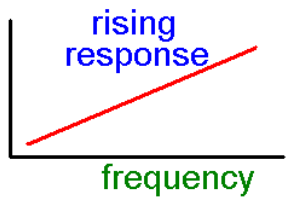
If we want a nice, squared-off pulse, we can feed the output of our RC differentiator into the "+" input of a comparator. We use a voltage divider to set the "-" input to a desired threshold value, and we get a rectangular pulse output.



An Opamp Differentiator

This circuit is a simple differentiator using the principle of the virtual-ground inverting amplifier. Note that I've flipped the opamp symbol over for convenience in drawing the diagram. Recall that an inverting amplifier's gain is the ratio of the feedback impedance to the input impedance. In this case, since the input component is a capacitor, the input impedance is inversely proportional to frequency. That means, for every doubling of input frequency, the gain doubles. This is technically referred to as a **rising response**. Note that this circuit inverts the polarity of the signal.

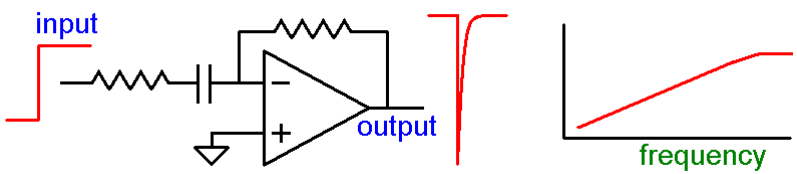




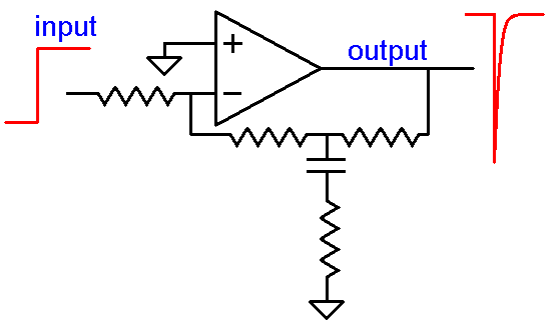
A Modified Differentiator for the Real World

True differentiation is often impractical or impossible. For example, an ideal square wave would have instantaneous rise and fall times, so the differentiated waveform would consist of positive and negative spikes of zero duration and infinite voltage. In the case of the inverting differentiator at lower left, the step input produces an output spike that will hit the negative rail. With regard to sinewave inputs, since response rises, doubling with every doubling of frequency, an extremely high input frequency would produce an extremely high output voltage. If we were differentiating, say, a 1kHz signal that had picked up a tiny bit of 1MHz RF noise, the differentiator would amplify that RF by a factor of 1,000X compared to the 1kHz signal. Undesirable, to say the least.

In practice, a differentiator circuit is built to have rising response *up to a point*. Our virtual-ground circuit would be modified with a series input resistor to limit the maximum gain. It would maintain a rising response as far up in frequency as needed to meet the requirements of a particular application, then level off.



Input impedance will change with frequency. If that's unacceptable, we can use the circuit below. As frequency rises, the capacitor increasingly shunts the negative feedback signal, meaning that the output voltage must increase to drive the inverting input to zero in normal virtual-ground fashion.



I may have confused you by first describing differentiation in terms of rate of change and then switching to a discussion of frequency response. But those two things are really two sides of the same coin. Obviously, frequency is related to time; it's the number of cycles per second, and rate of change of a voltage is also related to time. So we can write an equation that combines them:

$$\Delta V / \Delta T = f \times V_{PEAK} / 2\pi$$

Change in voltage over time equals the peak voltage (phasor length) times angular velocity. We can talk in terms of frequency instead of angular velocity if we divide the right side of the equation by 2π. A differentiator would reveal that a 1Hz, 1VRMS sine wave, which would peak at 1.4V, has a $\Delta V / \Delta T$ of:

$$\Delta V / \Delta T = 1 \times 1.4 / 6.3 = 0.225 V/S$$

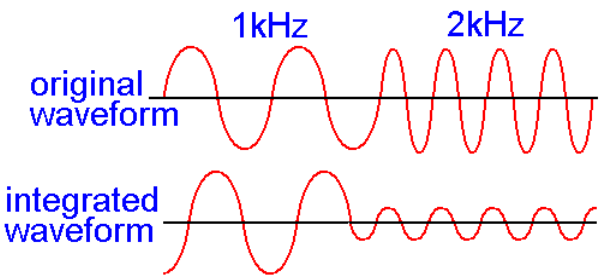
INTEGRATION

Integration is the opposite of differentiation. The output of an integrator is the input voltage averaged over time. Think of a bank account whose balance is the result of the accumulated deposits and withdrawals over a period of time.

As with the differentiator, we can't achieve theoretically perfect integration. To do so, the circuit would have to have been averaging the input voltage from the time it was manufactured.

Thus a real-world integrator averages the input voltage "for awhile," which in practical terms means long enough to find out what we need to know in a particular application.

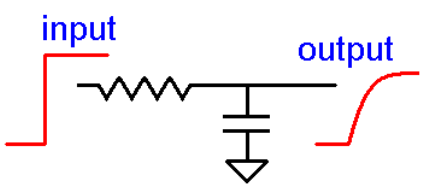
When a sine wave is fed into an integrator, the output voltage is inversely proportional to frequency, and the output phase lags the input.



The output signal of an integrator is referred to as the **integral** of the input signal.

The Simplest Integrator

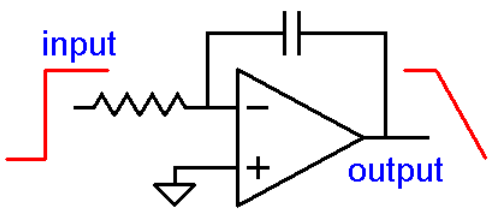
The simplest "almost-integrator" is a resistor and capacitor. The capacitor is made large enough to have a great deal of charge storage capacity. Voltage across the capacitor follows the input voltage,



but slowly. The RC time constant should be at least 5 times longer than the time period that we want the circuit to integrate. The load on its output should be very light, ideally an open circuit, because the series resistance is likely to be high; any load will upset the parameters.

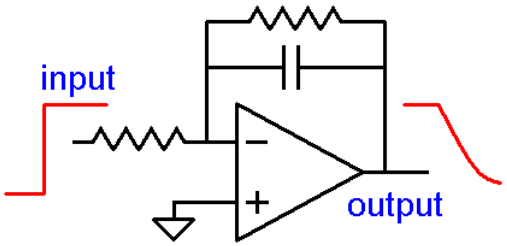
An Opamp Integrator

The opamp integrator is, again, based on an inverting virtual-ground circuit. When a DC voltage is applied to the input, the output slowly begins to go in the opposite direction as the capacitor charges. If the input signal is a sinewave, the output will be a sinewave that gets smaller as frequency rises.



If we were to build this circuit, we would find that the tiniest DC offset at the input would cause the output to slowly rise or fall until it hit one of the DC rails. In practice, we have to parallel the capacitor with a high-value resistor to provide a DC feedback path (diagram next page). This has the same effect as the input resistor we added to the differentiator circuit: as frequency falls, gain rises until it hits a limit defined by the ratio of the resistors.

FEEDBACK RESISTOR LIMITS INTEGRATION TIME



SQUARE-TRIANGLE CONVERSION



If a square wave is fed into an integrator, the output will be a triangle wave. The technique won't work if the frequency varies, because then the triangle wave voltage will vary. The f-like symbol is the mathematical sign for integration.

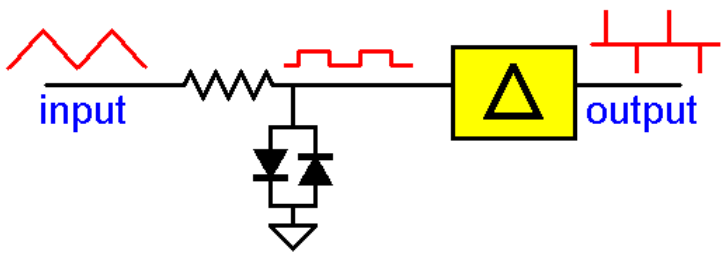


If a triangle wave is fed into a differentiator, the output will be a square wave. The triangle symbol is the mathematical sign for differentiation.

If the waveform duty cycle is not 50%, an integrator will convert a sawtooth to a rectangle wave. If we differentiate a triangle or sawtooth wave twice, we will get pulses that identify the points where the slope changes direction. This can be used to synchronize a sweep generator to an oscilloscope.

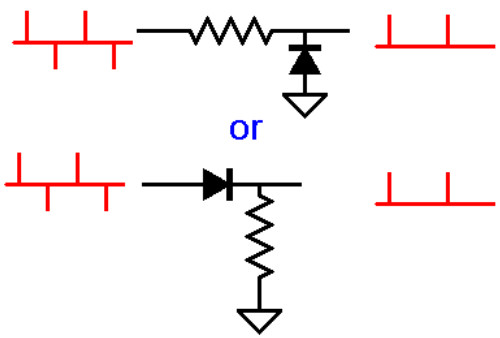
CLIPPERS

In most cases, clipping is something to be avoided. However, there are times when it's useful. A trapezoid wave isn't easy to generate, but we can derive it by feeding a triangle wave into a clipping circuit. If we feed a large sine or triangle wave into a clipping circuit, we get something close to a square wave. We can then differentiate this to derive the timing of the zero crossings. That can be used, for example, to trigger a thyristor switch during a zero crossing, avoiding the sharp wavefront that occurs when it's triggered during a positive or negative alternation.



Most audio signals contain voltage peaks far in excess of their average value. If we're interested in sound quality, we want to preserve those peaks. But if we're interested only in pushing the maximum signal through a communication channel – as in CB radio, for instance, where we're dealing with transmitter power of a few watts – we want to clip them off. This reduces sound quality, but we don't care as long as we can still understand what's being said.

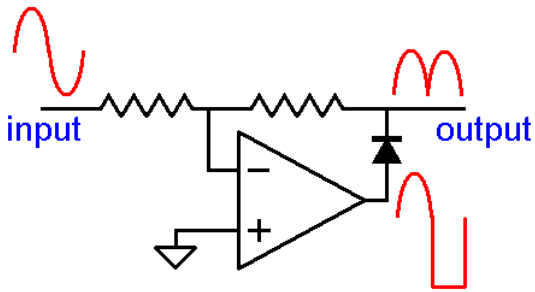
In an FM radio, we want to eliminate all amplitude variations in the RF signal. This is done by driving some of the amplifier stages far into clipping.



We can use a single diode to eliminate half of a waveform. Returning to the differentiated sweep generator example, we want a single oscilloscope-synchronizing pulse that corresponds to the beginning of the rising slope. The clipped, then differentiated waveform has two pulses, one positive, one negative. We can use a shunt diode to clip off the negative pulse, or a series diode to pass only the positive pulse.

PRECISION RECTIFIERS

One of the problems with rectifying a low AC voltage is the 0.7V drop across the diode. This can be improved by using a germanium diode, which only has a 0.3V drop. But there are many times when more precise rectification is needed.

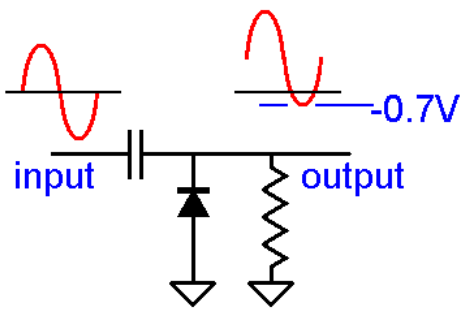


This circuit is a nearly perfect low-frequency rectifier. During positive signal alternations, the reverse-biased diode disconnects the opamp output. If the circuit is not loaded, i.e. if the output is fed into the input of a FET or MOSFET opamp, there will be no voltage drops across the resistors and output voltage will be the same as input voltage. During negative alternations, the diode is forward biased and the circuit acts as an inverting amplifier with unity gain.

On every positive signal alternation, the opamp output goes to the negative rail, comparator style. At higher frequencies, the time delays involved, plus diode capacitance, mess up the output waveform. These applications simply require more sophisticated rectifying circuits, which can be found in application notes published by opamp manufacturers.

CLAMPS

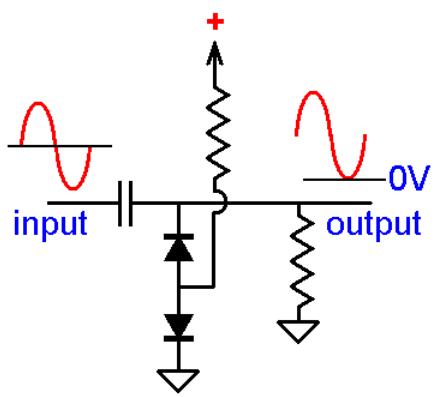
A clamp is a circuit that latches onto a part of a waveform and forces it to be a certain voltage, thereby setting the average DC level of the entire wave. The simplest clamp is a capacitor, high-value resistor and diode. The capacitor permits the DC voltage at the output to be different from the DC voltage at the input.



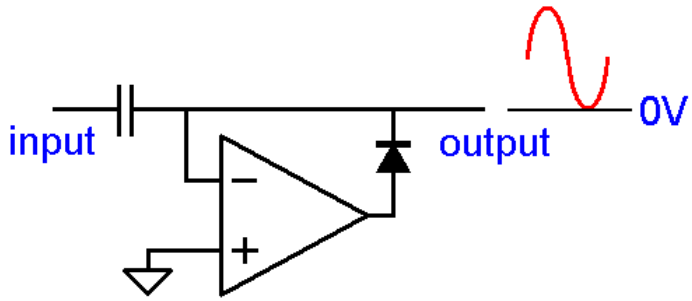
In the absence of signal, the resistor will discharge any DC at the output, insuring that the average DC voltage is zero. But if signal is applied, and if it tries to go more negative than -0.7V, the diode conducts, charging the capacitor. This shifts the average DC voltage so that, whatever the waveform may be, its most negative peak will be no lower than -0.7V.

Biased Clamps

What if we don't want the waveform to go 0.7V negative? We can bias the clamp by putting +0.7V on the diode.



If we want to get a little more sophisticated, we can use an opamp.

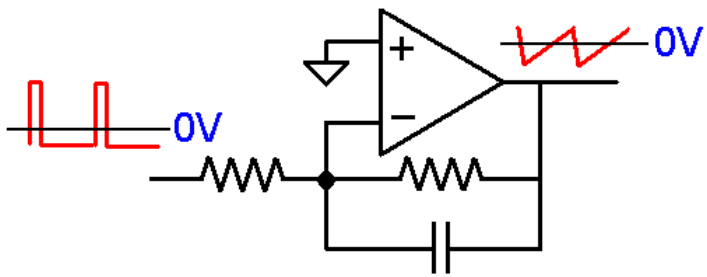


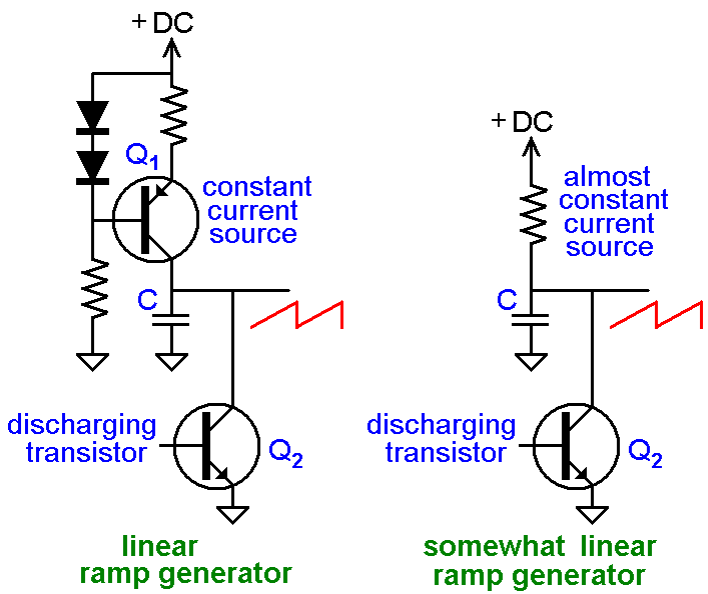
A common use for a clamp is to set the DC level of a baseband video signal. DC level is critical because it corresponds to overall picture brightness. The DC level is lost during transmission, but a clamp in the TV set restores it. A clamp also forms an RF-to-audio conversion circuit in some AM radios.

RAMP GENERATORS

You'll recall that a ramp is a smoothly rising or falling voltage. An RC charging circuit does not generate a linear ramp because charging current varies with time. To get a linear ramp, charging current must be constant.

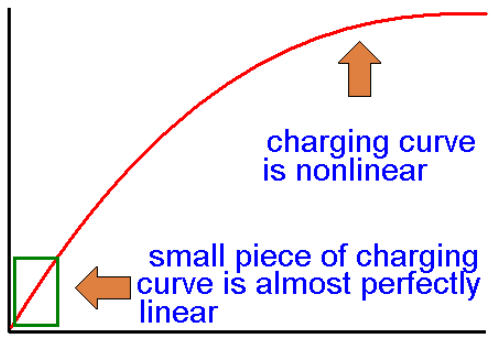
One way of obtaining a linear ramp is to feed a rectangular wave into an integrator. The positive- and negative-going portions of the output wave will be linear because the input of the integrator is a virtual ground, so current through the input resistor will stay constant over the course of a positive or negative alternation.





However, a ramp circuit application often involves specific thresholds which define the starting and ending points of the ramp. In these cases, an actual constant-current circuit works better. In the circuit below left, Q_1 is a constant-current source which charges capacitor C at a linear rate. When the ramp voltage reaches a preset threshold, a comparator (not shown) momentarily turns on Q_2, discharging the capacitor and starting a new ramp.

If you're not too concerned with linearity, a simple RC current source will give a fairly linear ramp if you only let the capacitor charge a little way. A small piece of a nonlinear curve is almost linear.



A circuit that changes a wave's shape usually changes its spectrum. In the next chapter, we're going to look at circuits where the whole idea is to change a signal's spectrum; that is, to treat some frequencies differently from other frequencies.

End of chapter 7

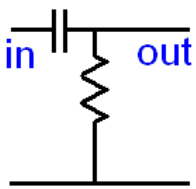
EM3 Chapter 8
FILTERS

Objectives of this chapter

- *define and discuss "filter"*
- *define and discuss "frequency response"*
- *discuss "turnover frequency" etc.*
- *describe circuits that would be used to achieve various frequency responses*
- *define and discuss selectivity and bandwidth*
- *discuss filter orders and attenuation slopes*
- *discuss L, T, π configurations*
- *regenerative filters*

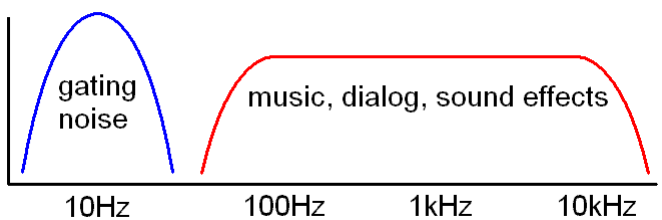
Electronic circuits do more than just convey signals; they also process them in various ways. As you know, real-world signals often contain energy at many frequencies. One of the most often-performed operations is to select some frequencies and reject others, as a strainer separates sand from gravel. A circuit that perform this function is called a **filter**.

A filter's frequency-versus-voltage characteristic (or any system's frequency-versus-voltage characteristic) is called its **frequency response**, and it's shown on a graph with frequency along the X-axis and voltage along the Y-axis. If a narrow band of frequencies is represented, the X-axis will be linear; if a wide band of frequencies is represented, it's more practical to use a logarithmic X-axis scale. Voltage on the Y-axis is usually measured in decibels. The vertical scale is linear, though the actual quantities are logarithmic; the linear-to-logarithmic conversion is inherent in the way decibels are calculated. If the filter has a wide frequency span where response is linear, that will be called "0dB" and everything else will be referenced to it. If you're a bit shaky on dB's, you may want to go back to chapter 1 and review.



Here's a very simple example of a filter. I described in EM2 how a capacitor can be used to block a DC voltage while at the same time passing an AC signal. But if we want to reduce low AC frequency energy too, we can just make the capacitor smaller than we otherwise would.

Here's one use for that: there's always a lot of very low-frequency noise in an optical movie sound track, down around 10Hz, called "gating noise." It isn't there because the director wanted it there, it's a byproduct of the process. It's not part of the program and we don't want to hear it.



So we make one or more of the coupling capacitors in the circuit smaller than we otherwise might, and that blocks the rumble while passing speech, music and sound effects.

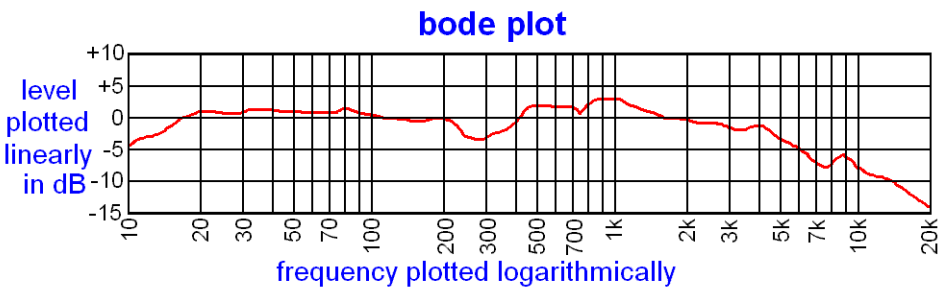
FILTER COMPONENTS AND FUNCTIONS

Both X_C and X_L change with frequency, so inductors and capacitors are the obvious components to use, along with resistors, when we want to favor some frequencies over others.

A filter can be thought of as a black box that accepts a signal at the input, modifies it in a frequency-selective way, and routes it to the output.



The way the input signal is modified by the filter is called a **transfer function**. This can be expressed in complex mathematical terms, but we're not going to bother. Instead, we're just going to think in terms of frequency response curves. In the graphs, signal strength will be plotted in decibels. A frequency vs. signal graph where the Y-axis is in decibels is called a **Bode** (BOW-dee) **plot**. Note that, since the frequency scale is logarithmic, there's no zero; the plot is simply truncated at however many cycles are required. And, since decibels are logarithmic, there's no absolute zero on the Y axis, though in the example below a zero reference (which is *not* zero volts) has been chosen, and other levels referenced to it.



TYPES OF FILTERS

There are dozens of kinds of filters. We're going to look at the following types:

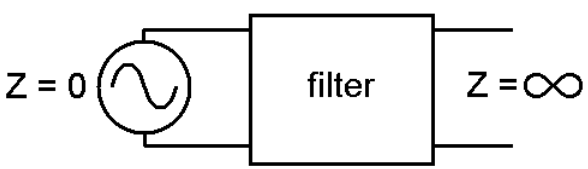
- RC & RL **low pass**
- RC & RL **high pass**
- RC & RL **band pass**
- RC & RL **band reject** or **band stop**
- **LC & RLC filters**
- **double-tuned** LC filters
- **all pass**
- **boost**

Filter names can also be properly written without a space: lowpass, highpass, bandpass, allpass, bandstop.

MATCHED AND NONMATCHED CIRCUITS

A filter's behavior is greatly influenced by what's connected to either end. Generally, a filter lives in one of two environments, impedance-matched and nonmatched.

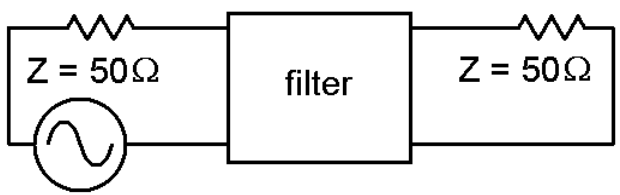
A **nonmatched** circuit is one where the signal source that precedes the filter has an output impedance that is so much lower than the filter's own impedance, and the load that follows has an impedance that's so much higher, that those impedances do not need to be considered.



Ideally, we'd be dealing with a stiff, zero-impedance voltage source and an open-circuit load. This simplifies the calculations because we don't have to consider anything beyond the filter itself. A filter inside a piece of electronic equipment is likely to be nonmatched.

When signals are sent over long wires, or are extremely high in frequency, this won't work. For reasons explained in the cabling chapter of EM2, it is often necessary for every input, output and transmission line in a system to have the same impedance. This is called an **impedance-matched** system. A filter in such a circuit must be designed to work with a source and load of this specific impedance.

Moreover, the filter itself should have this same impedance, so the source and load are connected to the impedance they are designed for. This is often difficult to do, as filter impedances tend to vary with frequency. It is usually sufficient for the filter to have the correct impedance over the band of frequencies that it passes unaltered, and only deviate greatly from this impedance at frequencies that are very much attenuated. In some cases, it is necessary to insert a resistor network before or after the filter to manage the impedances, at the cost of losing some signal level.

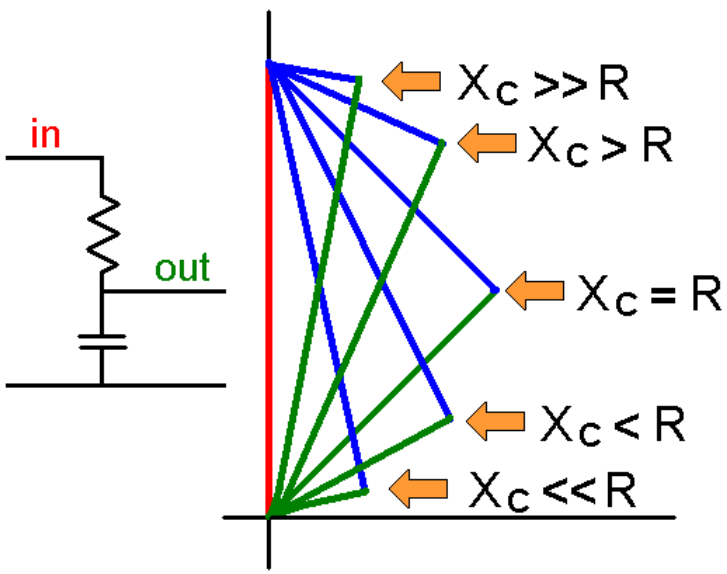


The easiest filters to understand and design are nonmatched RC and RL filters, so that's what we're going to start with.

VOLTAGE DIVIDERS AGAIN

In the illustration below, a series RC circuit has been set up to act as a voltage divider, with an input and an output. The red line represents the input voltage and the green lines represent the output voltage at various frequencies.

Let's say that the frequency where X_C = R is 1kHz. At this frequency, the voltage drops across R and C are both 0.707 x the input voltage. If you remember your dB's, or you've got the chart out, you know that a square-root-of-$^1/_2$ ratio



means the output is 3dB down from the input. The output phase lags the input by 45°.

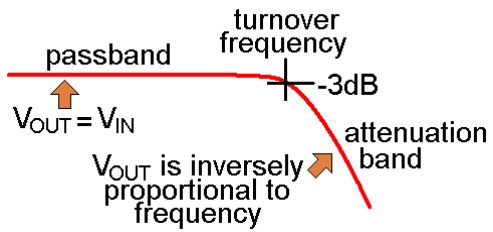
If the frequency rises, X_C falls. This changes the ratio; X_C is now smaller than R, and the voltage drop across C is proportionately lower. Since the voltage across the capacitor, the green line, is the output voltage, we can say the attenuation of the voltage divider increases as frequency rises. The phase difference also increases; the higher the frequency, the closer to 90° lag. If the frequency continues to rise, output voltage will continue to fall until, in theory, at frequency = infinity the output will be zero.

What about lower frequencies? Will the output voltage get larger? Yes, but not much. The output of an RC voltage divider is never higher than the input, and the input is 1V. At the lowest frequencies, XC will be so high that it doesn't have any effect; we'll just get 1V out. Also, the phase difference will approach 0°.

THE RC LOW-PASS FILTER

What we've done is create a circuit that passes signals up to 1kHz with little loss, but weakens signals above 1kHz. That is called a **low pass filter** because it passes low frequencies and attenuates high frequencies.

Reducing the amplitude of a signal is called **attenuation**. The low-frequency range, which the filter passes with little attenuation, is called the **passband**. The range of frequencies which it attenuates is called the **attenuation band**. The frequency where the passband and attenuation band meet is called the **transition frequency**, **turnover frequency**, or **cutoff frequency**.



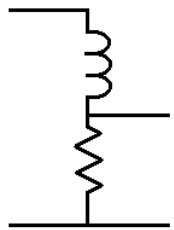
This junction point does not form a sharp "knee." Attenuation is already beginning in the upper part of the passband, and in the attenuation band does not reach its final slope immediately. Thus the choice of an "official" turnover frequency is somewhat arbitrary. By convention, and also for technical reasons, the turnover frequency is regarded as the point where response is 3dB down from the passband response. Among engineers and techs, "**3dB down point**" has the same kind of significance as "impedance"; it's the name of a concept as well as an aspect of circuit behavior.

- It's the dividing point between the passband and attenuation band.
- It's the frequency where the response falls to 0.707 times its passband value, and power falls to 0.5 times the passband value.
- It's the frequency where reactance of the capacitor equals resistance of the resistor.

An RC low pass filter operates by increasingly shunting the signal to ground as frequency rises, due to the falling reactance of the shunt capacitor.

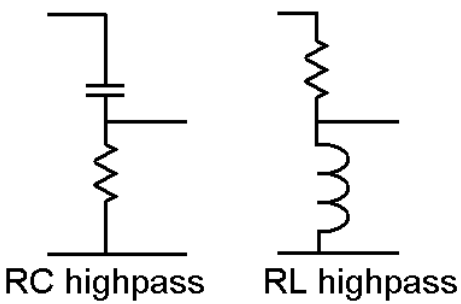
THE RL LOWPASS FILTER

In an RL low pass filter, the transfer function is the same but it's achieved in the opposite way. An RL low pass filter operates by increasingly blocking the signal as frequency rises, due to the rising reactance of the series inductor. The turnover is the frequency where $X_L = R$.



RC AND RL HIGH PASS FILTERS

If we flip the components around, we get a filter that passes high frequencies and attenuates low frequencies. The RL filter operates by shunting low frequencies as the reactance of the inductor falls.



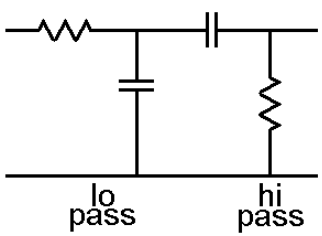
The RC filter operates by blocking low frequencies as the reactance of the capacitor rises. The frequency response curve has exactly the same shape as the high-pass curve, but flipped so the passband is to the right.

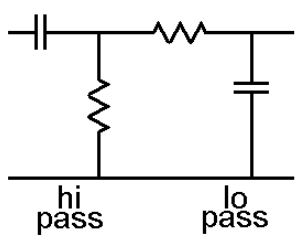
RC AND RL BAND PASS FILTERS

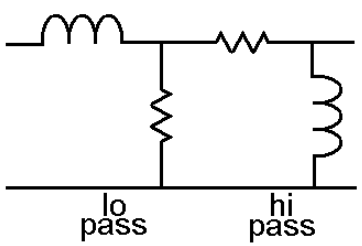
A band pass filter passes frequencies in a certain band and attenuates frequencies outside it.

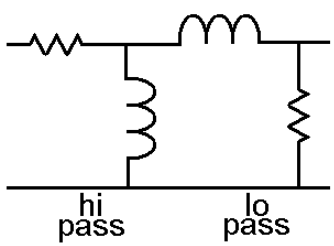


For a wide pass band, the most effective technique is to connect a low pass and a high pass filter in series. For a narrow pass band, a tuned filter is preferred, which we will discuss later.









TURNOVER FREQUENCY FORMULA

In an RC filter, the turnover frequency is the frequency where $X_C = R$. You remember how to find X_C at a particular frequency:

$$X_C = 10^6 / 2\pi fC$$

where f is frequency in Hz and C is capacitance in mfd.

To find the capacitance needed for a particular turnover frequency, first choose R, then substitute it for X_C in the formula.
To find the turnover frequency of an existing filter, rearrange the equation to isolate f on one side of the equal sign.

$$f = 10^6 / 2\pi RC, \text{ C in microfarads}$$

You can't depend on this formula to be completely accurate, because the source may not actually be zero ohms, the load may not be infinity, and the filter designer may have fudged the component values to take those issues into account. However, you'll get close enough to give you a good idea of whether the circuit is operating the way it's supposed to.

The formula for finding the turnover frequency in an RL filter is:

$$f = 10^3 R / 2\pi L$$ where L is in in mH

$$f = 10^6 R / 2\pi L$$ where L is in μH

SELECTING COMPONENT VALUES

A given turnover frequency can be obtained with many different component combinations. For example, a 1kHz high or low pass filter could be built with a 10kΩ resistor and a 0.016μfd capacitor, or 100kΩ and 0.0016μfd, or 1kΩ and 0.16μfd. So, do the specific values make any difference?

In a nonmatched system, there is usually a wide range of components that will give about the same results. One should stay away from very high values, especially if they're much higher than other values in the circuit. Even in a nonmatched system, most loads are not open circuits, and they may have shunt capacitance which could affect frequency response. Also, with very low-level signals, high-impedance circuits tend to be noisier. Up to a point, lower impedance is better. However, you have to make sure that the filter's input impedance is not low enough at any frequency that it overloads the signal source, causing distortion or some other malfunction. In general, the ideal values are those that are higher than the source and lower than the load by wide enough margins that source and load impedances are unimportant in the overall scheme of things.

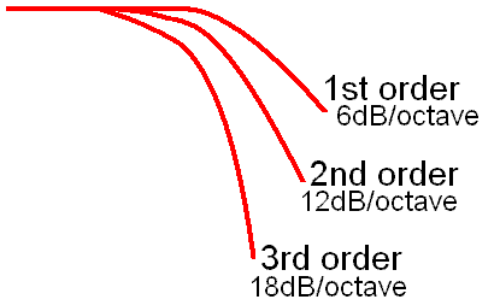
In a matched system, you simply have to take source and load impedances as givens. If they are mostly resistive, use standard Ohm's Law calculations to find series and parallel values. A lot more could be said, but I don't think it's worth your effort to study at this point, because the filters used in matched systems are usually not RC or RL but LC filters, which we'll talk about later.

SLOPE AND ORDER

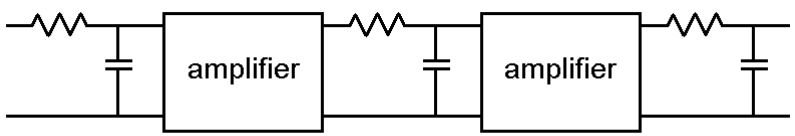
The rate at which response falls off in the attenuation band is called the **slope**. In a simple 2-component RC or RL filter, voltage varies directly with frequency in a high-pass filter; every time the frequency is halved, output voltage is halved. In a low-pass filter, voltage varies inversely with frequency.

This 1:1 correspondence results in a **6dB per octave** slope. As you know, 6dB is a 2:1 voltage ratio and an octave is a 2:1 frequency ratio. This is the natural slope of a 2-component RC or RL filter. A filter with one reactive component and a 6dB/O slope is called a **first-order** filter.

If we want attenuation to increase at a faster rate, we can connect filters in series. This is called **cascading**. Two cascaded first-order filters will result in a **second-order** filter with a **12dB/O** slope; three cascaded filters will result in a **third-order** filter with an **18dB/O** slope and so on.



There's a potential problem with cascading, because the first filter does not see an infinite load impedance and the second filter does not see a stiff source. The problem can be eased by choosing the values in such a way that each successive filter is much higher impedance than the one before. However, the best method is to put a unity-gain (1X) amplifier between the filter stages to isolate them.



3rd-order low pass filter

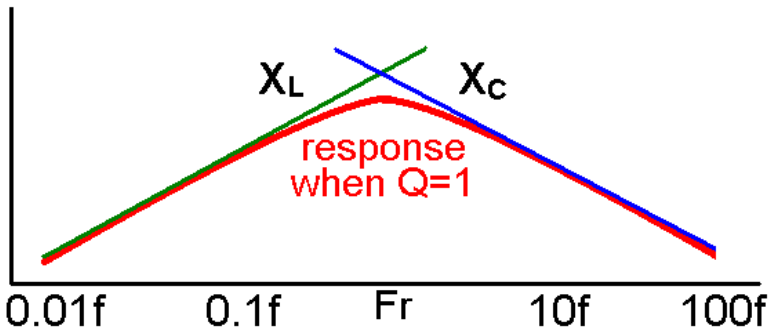
We've still got a problem. A 1st-order filter is 3dB down at the turnover frequency. If we cascade three filters, the resulting filter will be 9dB down at the turnover frequency. 9dB down is serious attenuation.

If we want the response to be fairly linear until the beginning of the attenuation band, we need to come up with a circuit that gives us a 2nd-order or 3rd-order slope, but is only 3dB down at the turnover frequency. There are two ways of doing this. One is to build an **active filter**, which uses the gain of an amplifier to alter the curve near cutoff. We'll look at active filters later in this chapter.

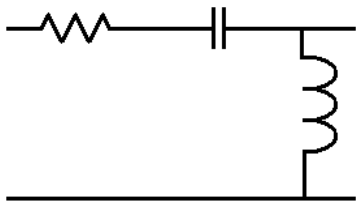
RLC L0W AND HIGH PASS FILTERS

The other way to get a sharp knee is to build an **RLC filter**, which combines inductive and capacitive reactances with resistance. Since an RLC filter has two reactive components, it gives a double knockdown to frequencies in the attenuation band. In a low pass filter, the inductor blocks them and the capacitor shunts them. Thus an RLC high- or low-pass filter is automatically a 2nd-order filter with a 12dB/O slope.

In an RLC high- or low-pass filter, the turnover frequency is approximately equal to the resonant frequency of the LC combination, which is the frequency where $X_L = X_C$.

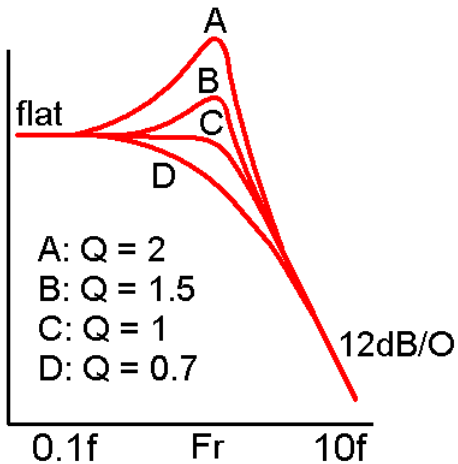


You've seen the illustration above, showing that the response of a parallel RLC tank with a Q of 1. If we rearrange the components to form a low-pass filter, we'll get a curve that's 3dB down at cutoff and 12dB/octave thereafter. Already we've improved on the double first-order filter, which is down 6dB at the turnover frequency.



But we can do more than that. By controlling the Q of the filter, we can get a sharp knee, a rounded knee, or a peak, as shown above right. At frequencies far below f, the response is the same for all values of Q – that is, it's flat – and at frequencies far above Fr it falls at 12dB/octave. Q only affects response at and near Fr. Also note that, when the Q is higher than 1, the attenuation slope is initially greater than 12dB/O.

The RLC high-pass filter characteristic is an exact mirror image of the low-pass filter. The response is flat far above f, changes at 12dB/octave far below f, and is controlled by Q at and near Fr.



LC TURNOVER FREQUENCY FORMULA

Since the turnover frequency is the LC resonant frequency, it is found using the same formula used to find an LC resonance:

$$f = 1 / 2\pi(L^2 + C^2)^{0.5}$$

or

$$1.59/ (LC)^{0.5}$$

Inductance is in Henrys and capacitance in Farads, so unless you're dealing with a very unusual filter, some decimal shifting will need to be done.

When there's a lot of resistance in the circuit – that is, when Q is low – the formula becomes slightly imprecise. Most of the time, we don't have to care, because when Q is low the "knee" between passband and attenuation band is rounded and the transition is gradual anyway.

DAMPING

As you remember from EM2, damping is the inverse of Q:

$$D = 1 / Q$$
$$Q = 1 / D$$

The higher the Q, the lower the damping.

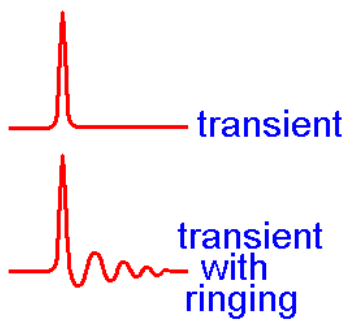
Techs are in the habit of talking in terms of damping more often than Q when dealing with high pass or low pass filters, because then a high Q is something we usually don't want. In most cases, a damping factor around 1 is desired. That gives us just enough resonant behavior to provide a flat response almost to the turnover frequency, with just the tiniest peak. An RLC circuit with a damping factor of 1 is said to be **critically damped**.

RESPONSE SHAPES NEAR CUTOFF

A gentle transition between passband and attenuation band, and a mild 6dB/oct rolloff, is often the correct filter characteristic, which is what we get from a first-order RC or RL filter.

On the other hand, there are times when we need an aggressive filter. An ideal filter would have the following characteristics:

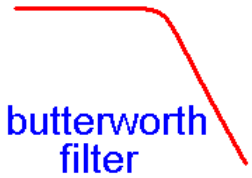
- completely flat response in the passband
- very fast rolloff in the attenuation band
- little or no ringing on transients



By **ringing** I mean the way a resonant circuit oscillates for a while every time it's hit with a transient. This is an often-overlooked factor in filter design. Unfortunately, the three ideals are, to an extent, mutually exclusive. We can have completely flat response or a super-fast rolloff or lack of ringing, but not two of those and certainly not all three. Different engineers have settled on different compromises, and several of them have names.

The Butterworth Filter

A Butterworth filter is optimized for flattest frequency response. The response is perfectly linear until nearly the turnover frequency, where it falls to -3dB.



A Butterworth filter rings on transients, and the faster the rolloff, the worse the ringing.

The Chebyshev Filter

A Chebyshev filter is optimized for very fast attenuation outside the passband. The cost is a slight **ripple**, meaning peaks and dips in the response, inside the passband.



The designer can choose how much ripple to permit; the more ripple permitted, the faster the rolloff. A Chebyshev filter with zero ripple is essentially a Butterworth filter.

Chebyshev filters have more of a ringing problem than Butterworth filters.

The Bessel (sometimes called Thompson) Filter

A Bessel filter is optimized for minimum phase shift in the passband, minimum ringing on transients, and the same **group delay** between input and output at all passband frequencies. It's an attempt to do the necessary filtering with the least possible alteration of signal waveforms. It achieves these virtues at the cost of a transition between passband and stopband that begins well before the turnover frequency; that is, it has a sloping “knee.” Thus it does not offer flat response up to the turnover frequency, as a Butterworth filter does.



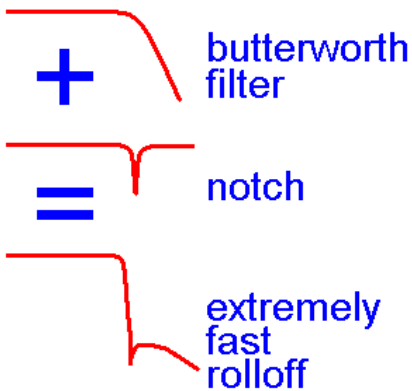
Far from the turnover frequency and well into the attenuation band, the attenuation slope for all three filters is the same. Their differences are in how they behave near the turnover frequency.

FILTERS INCORPORATING A NOTCH

If we cascade more filter stages, we get a faster and faster rolloff, but we also increase any problems near the turnover frequency and we increase ringing.

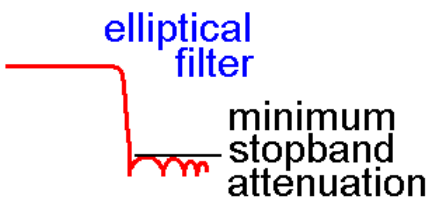
There's another way to get a fast rolloff, and that is to choose a frequency just outside the passband and put a filter that provides great attenuation, just at and on either side of that frequency. In other words, it puts a **notch** in the frequency response. That will give us almost total attenuation at frequencies just above the passband.

At frequencies farther removed from the turnover frequency, the notch filter will no longer be effective, but by then the main filter's regular rolloff will have reached a high degree of attenuation.



The Elliptic or Cauer Filter

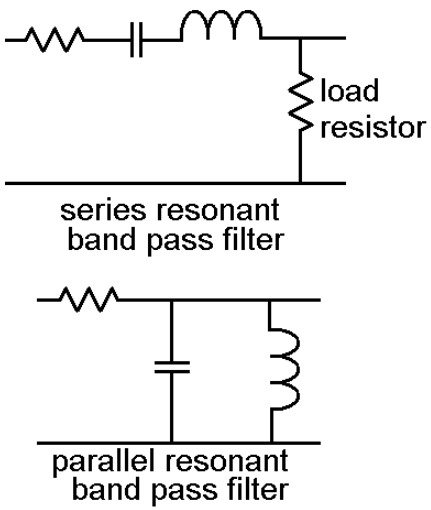
A more elegant and better performing filter with more than one notch outside the passband is the **elliptic** filter, often called a **Cauer** filter. In many applications, it is not important for attenuation outside the passband to keep increasing as the frequency gets farther from the turnover frequency. Instead, we often just want attenuation to hit a certain benchmark, a certain minimum, just outside the passband. From there on out, we don't care if the attenuation continues to increase, only that it never decreases so as to rise above the specified limit at any frequency in the stopband. That's the characteristic of an elliptic filter. Its frequency response in the attenuation band, instead of continuing to fall, is flat with some ripple, like the passband. However, it reaches that attenuation just outside the passband and is never poorer than that at any frequency.



A filter whose attenuation slope is very steep is often called a **brick wall** filter.

RLC BANDPASS FILTERS

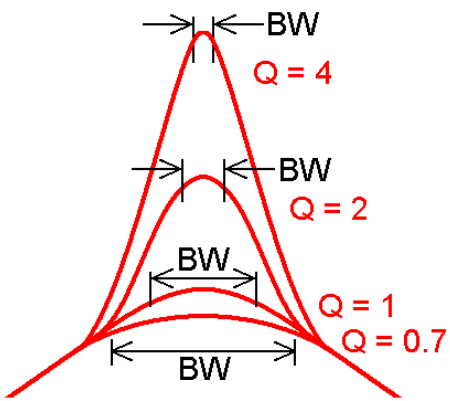
A 2nd-order RLC bandpass filter can be either series or parallel. It consists of a capacitor and inductor in a resonant circuit, and a resistance to control Q. The illustration shows one example of each type; you will run into others.



Bear in mind, too, that source and load impedances may play a part, especially with a filter used in a matched-impedance system.

In a bandpass filter with one inductor and one capacitor, the slopes on either side of Fr are 6dB/octave. This is still a 2nd-order filter, but instead of getting a 12dB slope, we get two 6dB slopes. At and near Fr, response depends on Q.

BANDWIDTH



The bandwidth of a band-pass filter is defined by the points at which response is 3dB down, compared to the response at Fr. The higher the Q, the narrower the bandwidth. Bandwidth can be stated as a frequency range, or as a percentage of Fr. For example, a 1kHz bandpass filter with 3dB down points at 950Hz and 1,050Hz would have a bandwidth of 100Hz, or 10%. Formula:

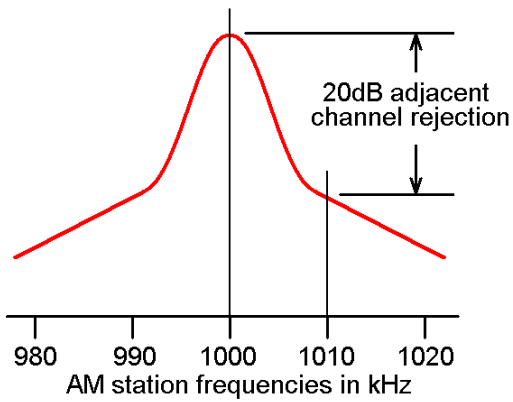
$$Q = 2\,Fr / B$$

where Fr is the resonant frequency in Hz and B is the frequency span between 3dB down points in Hz.

SELECTIVITY

Selectivity is a measure applied to radios and other devices that must pass one frequency and block another. For example, AM stations are assigned frequency **channels** 10kHz apart on the dial. A tuned circuit must pass one station's signal and attenuate adjacent signals as much as possible. The amount of attenuation of unwanted stations is called the **selectivity** and it's measured in decibels. Selectivity may be separately specified for frequencies one, two and three channels away from the selected channel.

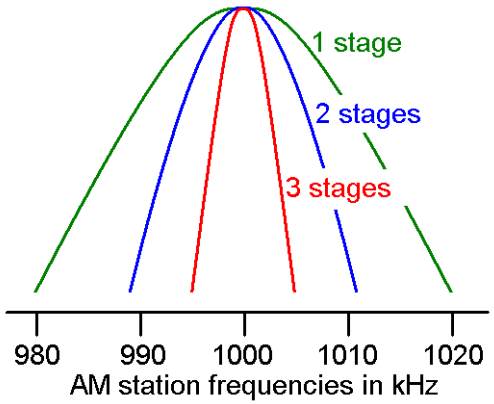
The illustration shows a band pass filter tuned to an AM station at 1MHz. Adjacent channel selectivity is 20dB.



CASCADED BANDPASS FILTERS

A filter with 20dB adjacent-channel selectivity wouldn't be anywhere near good enough for a real radio. That's why real radios send the signal through several cascaded filters, isolated from each other by amplifiers.

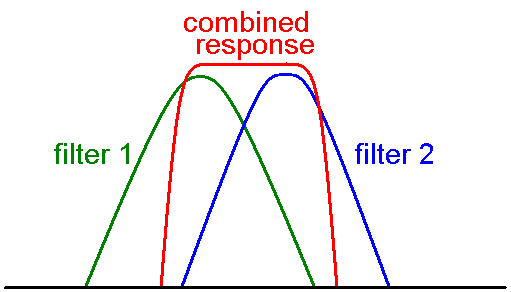
The illustration shows three cascaded filters that each have 20dB adjacent-channel selectivity. The total is 60dB, which is an acceptable figure for a real-world radio.



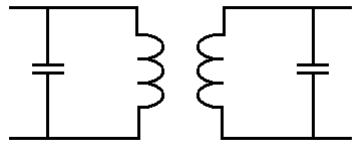
This is how radios were made in the 1920s. Radios have advanced a lot since then. I'm not going to go into great detail because we have all of EM6 to talk about radios, but I do want to mention something you'll run into in other electronic fields as well.

It's often desirable to have a passband response that's linear over a defined band of frequencies, then drops off sharply outside this band. With cascaded filters, we have an opportunity to do that.

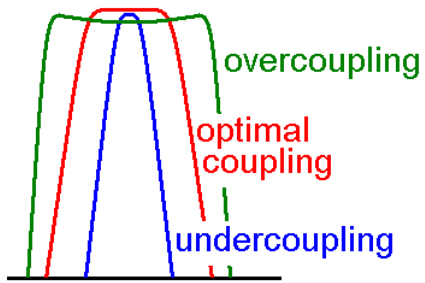
One method is to stagger the tuning slightly, so the f of one filter is a bit low and the f of a second is a bit high. If designed carefully, the combined response will have a flat top and steep slopes.



The other method is to make use of the mutual inductance of a transformer. We start by building two independent tank circuits tuned to the same frequency. Then we start bringing the inductors close to each other, so some mutual inductance develops. This will create a **double-tuned** transformer with a very sharply tuned bandpass.



If we keep bringing the inductors toward each other, so as to increase mutual inductance, we will reach a point where the coupling is maximized. Beyond that point, the passband will spread out, forming a flat-topped shape. If we continue to increase the coupling, we'll find that the response at Fr actually declines a bit, resulting in a wide passband with a slight dip, often called a “saddle-shaped” response. This is not necessarily a mistake; somewhere else in the system there may be a tuned circuit with a rounded curve and the combination will result in a flat top.

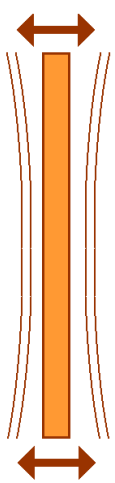


CERAMIC AND CRYSTAL FILTERS

When we studied the inductor and capacitor in EM2, we noted that an LC combination resonates like a mass and a spring. They're no reason a resonant circuit can't be an actual mass and spring if we have a way of converting electricity to motion and back again.

A quartz crystal can give us everything we need to do this. To review from chapter 6, quartz flexes in response to a voltage – many cellphone speakers are quartz crystals – and it generates a voltage in response to flexing. This is called the **piezoelectric** effect. We've seen how this can be put to use in an oscillator. But a crystal can also be used as a filter.

A mechanical chime or doorbell is essentially a bar or rod that flexes. Its mass provides the necessary mass and its stiffness provides the necessary spring to form a resonant system. A quartz crystal is a miniature chime. It resonates mechanically the way an LC circuit resonates electrically, and its piezoelectric nature gives us a way to convert electrical energy to mechanical and back again.



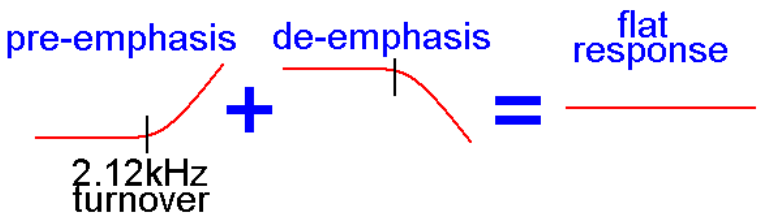
It is possible to make a filter using a quartz crystal that is extremely accurate, has very steep attenuation slopes, consumes little power, and is cheap. A variation is the **ceramic filter**, which combines a ceramic mass with quartz crystal. Very complex RLC filters can often be replaced with equivalent ceramic filters at a fraction of the cost and complexity.

One very common application is in radio receivers, which we'll study in great detail in EM6. We've already looked at the flat-topped response of a double-tuned transformer. A ceramic filter can often provide the same kind of response much more simply.

EQUALIZATION

Sometimes, instead of changing a system's frequency response, we use a filter to counteract a change that has occurred elsewhere in the system. For example, if high frequencies have been attenuated, which is something long cables tend to do, we can send the signal through a filter that attenuates everything *except* high frequencies, thus restoring the original frequency balance. This process is called **equalization**. The most sophisticated example at the moment is probably found in digital data transmission over a cable. In the highest-frequency examples, when the system is first connected, it uses a standard test signal to “learn” the cable's high-frequency loss characteristic, and automatically provides equalization to correct the overall response.

In some cases, a filter is used to deliberately boost high frequencies, then another filter is used to attenuate them back to where they were. This is called **pre-emphasis** and **de-emphasis**. One well-known example is found in FM radio. Most of the noise or "hiss" in FM is in the higher audio frequencies. This noise can be greatly reduced by reducing the high-frequency response of the receiver – turning down the treble knob, for example – but that also attenuates high frequencies in speech and music, which we don't want to happen.



So, to keep from damaging the sound, high frequencies are boosted or **pre-emphasized** at the transmitter according to a standard formula that's the same for all radio stations. In the receiver, high frequencies are attenuated or **de-emphasized** by an inverse formula that's the same for all receivers. The result is an overall flat response, but with reduced noise.

Over time, "equalization" has come to mean not just correction, but any intentional alteration of frequency response. For example, in voice communication applications like CB and police radios, it is customary to emphasize frequencies between 1kHz and 3kHz. This is not a restoration but an intentional alteration of response. The "naturalness" of speech is lost, but clarity is improved.

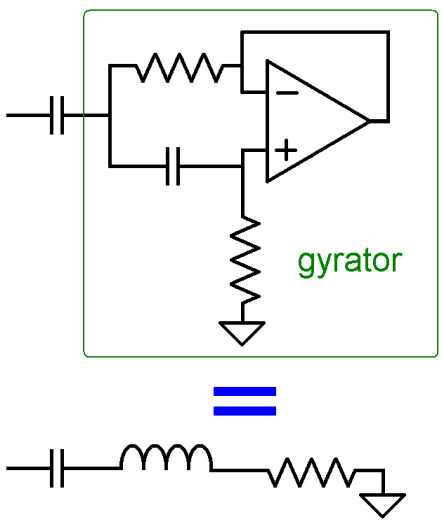
Up to this point, we've been looking at passive filters. None of them include an amplifier or require a DC power source.

ACTIVE FILTERS

An active filter is a one that uses an active device, and relies on it for its operation. Most active filters are based on opamps today, so that's what we'll concentrate on.

THE GYRATOR

The most basic active filter is one that uses active circuitry in place of the inductor in an LC filter, called a **gyrator**. Adding a series capacitor as shown creates the equivalent of a series-resonant LC circuit.

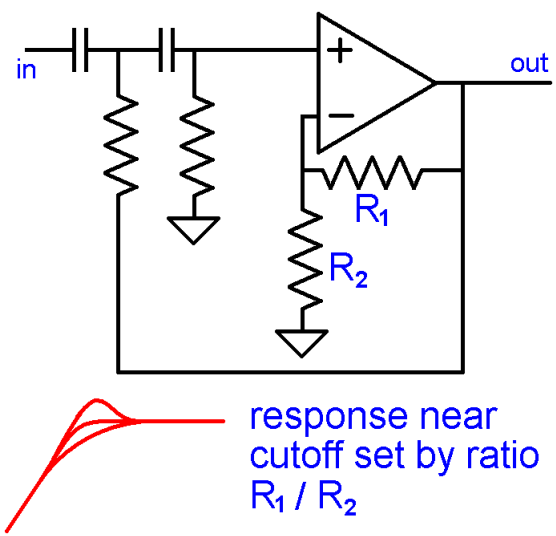


With suitable component choices, this circuit could resonate at 0.1Hz, which would be a very difficult thing to achieve with an LC circuit.

Placing this circuit across a signal line would cause a response dip or notch at the resonant frequency. On the other hand, it could be connected into the negative feedback loop of an amplifier in such a way as to cause a peak in its response.

SALLEN-KEY FILTERS

Sallen-Key filters were developed by a couple of guys named Sallen and Key, as if you couldn't have guessed. They use feedback to actively alter the response of RC filters. Sallen-Key filters cover a lot of territory, so I'll just show you one example. The illustration is a 2nd order RC high pass filter with a 12dB/octave rolloff.



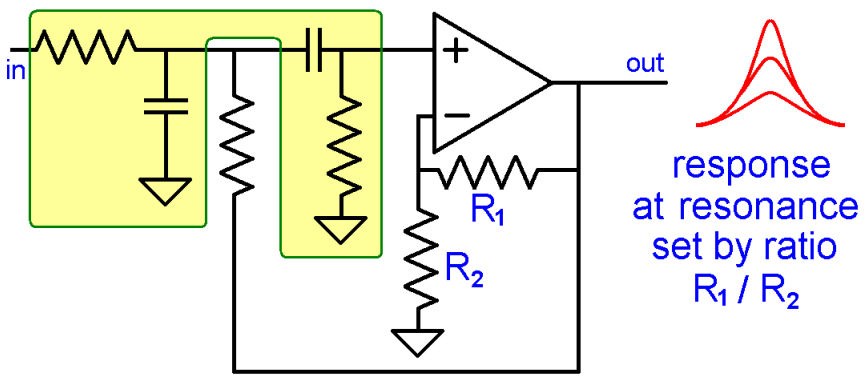
Normally, such a filter would be 6dB down at the turnover frequency f. But a positive feedback loop applies just enough boost at frequencies near f to lift the response, giving us an attenuation at f of 3dB.

Or not. Response at the turnover frequency depends on Q, which depends on gain, which is set by the negative feedback circuit, R_1 and R_2 in the diagram. A voltage follower, with the output directly connected to the inverting input, results in the lowest gain (unity) and a Q of 0.5. If $R_1 = R_2$, Q will be 1. Raising R_1 creates a peak, and when $R_1 = 2 \times R_2$, the circuit oscillates. By combining several Sallen-Key filters with carefully selected Q values, we can get a Butterworth, Bessel or Chebyshev filter with any rolloff rate we want.

MULTIPLE FEEDBACK BANDPASS

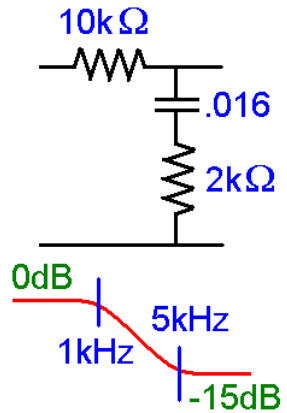
Similar techniques can give us a bandpass filter that's as sharply tuned as any passive LC filter. At higher radio frequencies there may be no point, since LC filters work really well and are passive (require no power). However, up to maybe 100kHz, an active, **multiple feedback bandpass filter** may be a more practical choice. It requires an opamp, thus a power supply, but it's usually part of equipment where there's a power supply available anyway.

In the circuit below, the yellow shaded parts amount to a standard RC lowpass followed by a standard RC highpass. If we wanted a wide bandpass, those circuits are what we'd use. But to get a narrow, resonant characteristic, we add positive feedback as before. Here again, response is determined by the R_1/R_2 ratio; when $R_1 = 2R_2$ the Q is infinite and the circuit begins to oscillate.



There are other kinds of active filter, like **elliptical**, **state-variable** and **constant-Q** filters, which are a bit more complicated. They operate on the same principle: a combination of positive and negative feedback to achieve response curves that couldn't be achieved by RC circuits alone. If you really want to get into this, you should pick up a copy of The Active Filter Cookbook by Don Lancaster. As a tech, you may never design a filter, but Lancaster's book will familiarize you with what's out there so you'll always know what you're looking at. It's the most practical active filter guide, and has been for decades.

SHELVING FILTERS



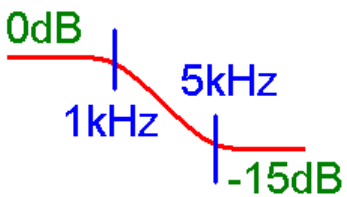
In some cases we want a filter's curve to flatten out again when it has reached a certain degree of attenuation. This is called a **shelving** filter and it's obtained by adding a resistor to create a second turnover point. The example at right is a lowpass filter that starts attenuating at 1kHz and flattens out again at 5kHz. At frequencies well below 1kHz, C is essentially open, so attenuation is approximately zero. At frequencies well above 5kHz, C is essentially a short, so the circuit is a 6:1 resistive voltage divider. That works out to a little over 15dB.

One common example of a shelving filter is a treble control in a stereo system. We may want to attenuate high frequencies beginning at, say, 2kHz, but we don't want to lose them entirely. So, a treble control typically starts attenuating at 2kHz and flattens out again around 8kHz. A shelving bass control does the same thing at low frequencies.

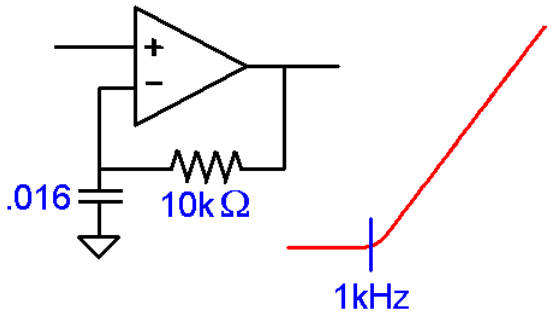
BOOST FILTERS

We've been talking about filters to attenuate certain frequency bands, but we often want to do the opposite – we often want to boost a range of frequencies.

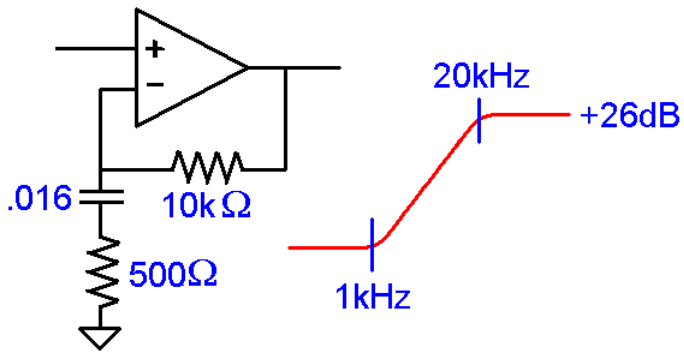
A passive RC circuit cannot boost anything; all it can do is attenuate. But it can give us the same effect as a boost by attenuating everything *but* the frequency range we want boosted. For example, we think of the shelving filter on the previous page as attenuating frequencies above 1kHz. But it could just as properly be thought of as boosting frequencies below 5kHz, relative to higher frequencies, at the cost of an overall 15dB loss of level. If we added an amplifier with 15dB of gain, we'd have a real low-boost filter.



But then, if we're going to have an amplifier, we can get boost by fiddling with the feedback circuit. For example, the amplifier above right has been designed to boost high frequencies. At 1kHz, the reactance of the capacitor is 10k Ohms. It rises at low frequencies, becoming essentially an open circuit, which results in a unity-gain voltage follower at low frequencies. But at progressively higher frequencies, its reactance falls, creating an ever-increasing voltage divider ratio and ever-increasing high frequency boost.

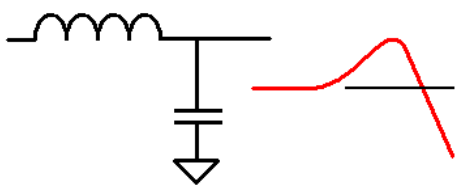


A circuit that exhibits ever-increasing boost is impractical for more reasons than I want to go into here. Take my word for it, we need to limit the amount of boost. That's where shelving comes in. The circuit below has been turned into a shelving boost filter, with the high frequency shelf at 20kHz and maximum boost at higher frequencies of just a tad over +26dB. At frequencies above 20kHz, the capacitor's reactance decreases to the point where the



feedback circuit is a straight 21:1 resistive divider. As a general rule, boost filters are shelving filters.

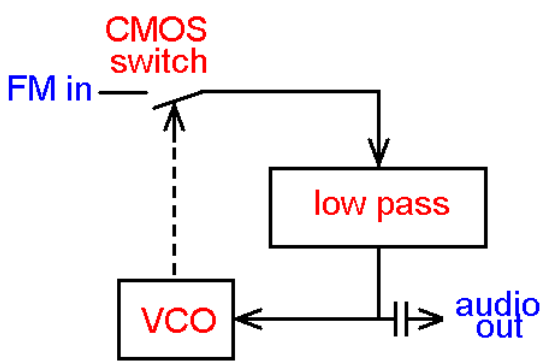
If we want a circuit that boosts up to a point and then falls off, we can use an active filter such as a Sallen-Key circuit, tuned so as to create a peak in its response. Alternately, we can use an LC or RLC filter, which is the only kind of passive circuit whose output voltage can be higher than its input voltage over a limited frequency range.



PHASE LOCKED LOOP FILTERS

In chapter 5, we looked at the phase locked loop (PLL), a circuit that uses feedback to control a voltage-controlled oscillator (VCO) so it always runs at the same frequency as the input signal. Usually, the oscillator control doesn't change on a cycle-by-cycle basis. Instead, instantaneous frequency errors are averaged by a low pass filter, so the VCO only responds to comparatively long-term frequency shifts.

Although this is not a filter function in the usual sense, a PLL can act like a filter. For example, this diagram is part of an FM radio. The input signal is around 10MHz; the low pass filter's turnover frequency is 50kHz. The filter output is a DC signal that varies in response to signal frequency deviations, feeding back a correction voltage that changes the VCO frequency to match.



But hey – that correction voltage can change 50,000 times a second. That's AC, dude. If we put a blocking capacitor in the circuit to remove the DC component, we get audio out.

We're going to talk a lot about phase locked loops in EM6, because they're used all over the place in RF and wireless.

FERRITE INTERFERENCE FILTERS

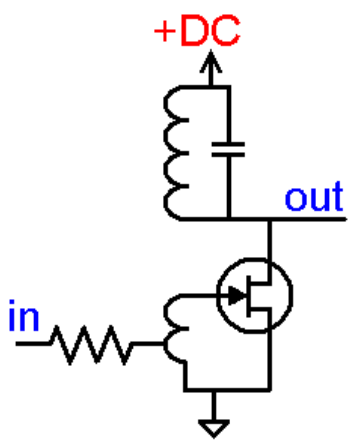
You know that lump on the power cord going to your computer is there to reduce radiated interference, right? But do you know what's inside it? A ferrite tube around the wires, that's what.



Even a straight wire has inductance, which is related to the electrical field that we learned, way back in EM1, surrounds all current-carrying wires. In most modern electronic equipment, the power supply has circuitry that switches high power at a high frequency, e.g. 100kHz. A significant amount of that gets back into the power cord, and if it were not blocked, it would radiate all over the place. The ferrite cylinder surrounding the power cord greatly increases the wires' inductance at that point. As you know, inductors block high frequencies. Problem solved. You'll find ferrite donuts inside equipment too, where there's a wire that must carry power from one place to another but should not carry the high frequency hash that may come with it.

REGENERATIVE FILTERS

I mentioned in the oscillator chapter that if we take a tuned oscillator and reduce the positive feedback so it's just below the point of oscillation, we'll have a very high-Q, high-gain filter at the resonant frequency. Here's one example: an Armstrong / Meissner filter where the coupling between the tickler coil and the tank coil is made weak enough that the circuit doesn't oscillate. Despite its simplicity, this is a practical, workable circuit.



DIGITAL FILTERS

Digital technology has given us new kinds of filters. Digital's way of handling analog signals is to chop them up and convert them into numbers. Once they are in number form, we can use mathematical processes to perform operations that would be complicated or impossible with analog filters. For example, we can make sharp-cutoff filters with little or no ringing. Since you have to learn about digital before you can deal with these filters, we're going to wait to talk about them until EM4.

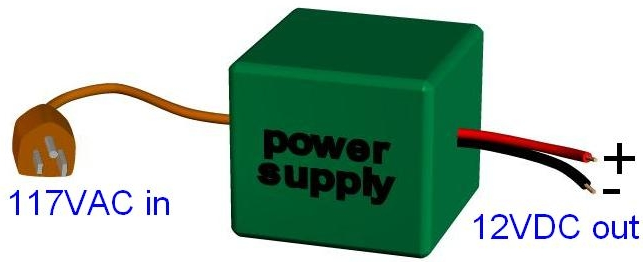
End of Chapter 8

EM3 Chapter 9
POWER SUPPLIES

Objectives of this chapter:

- *define "power supply"*
- *name the four parts of a linear power supply and what they do*
- *discuss half wave, full wave and bridge rectification*
- *explain the characteristics of a capacitor-input filter*
- *explain the characteristics of an inductor-input filter*
- *define "swinging choke"*
- *discuss resonant transformers*
- *explain power supply decoupling and successive filtering*
- *discuss shunt and series regulation and LDO regulators*
- *define "servo loop"*
- *describe how switching regulators work*
- *discuss boost, buck and boost/buck regulators*
- *describe the design of a modern switching power supply*
- *define and discuss "synchronous rectification"*
- *discuss square wave, trapezoid wave and sine wave inverters*
- *define and discuss "uninterruptible power supply"*
- *discuss variable frequency AC power supplies*

Unless a piece of equipment is battery-powered, it's almost certain to get its operating power from the AC power line. Even if battery-powered, the batteries will probably have to be recharged from the AC power line.

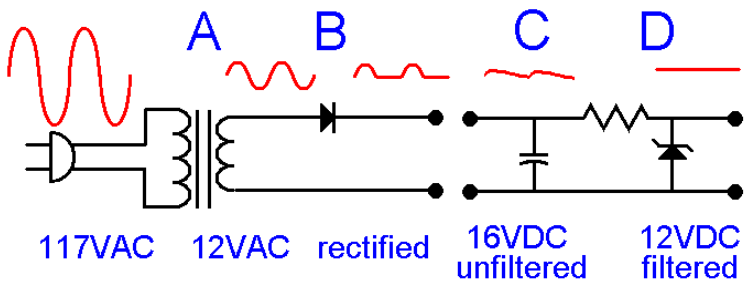


Our power generation and distribution system was originally designed to power lamps and motors, not computers and video screens. 117VAC is good for powering electrical equipment, but most electronic equipment requires a DC power supply at some voltage other than 117V. So we need to have a way of converting 117VAC to DC at a different voltage. That's what a **power supply** does.

We're going to assume we're dealing with supplies designed to put out a positive DC voltage with respect to equipment ground. A negative power supply can be built by reversing the polarities of the components. Most power supplies fall into one of two categories: **linear** (analog) or **switching**.

LINEAR POWER SUPPLIES

A typical small power supply has four parts, as shown in the diagram below.

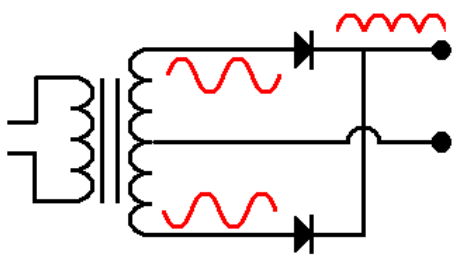


A) 117VAC is converted to a different AC voltage that's more suited to the circuit being powered. This is usually done by a transformer, which also isolates the circuitry from the AC power line.
B) AC is converted to DC. This is done by inserting a rectifier in the line. The resulting DC is “lumpy” because it follows the shape of the AC half-waves.
C) A capacitor is connected across the line (connect the dots in the diagram) to turn lumpy DC into reasonably smooth DC.
D) If “reasonably smooth” isn't good enough, the DC is further processed by a resistor/zener regulator to produce tightly regulated, AC-free DC voltage.

It's common to find this circuit supplying DC to very low-power electronic gear. However, there are many ways to improve it. We can do a better job of rectifying, we can filter more effectively, we can regulate more efficiently. We covered transformers in EM2, so let's start at “B,” the rectification stage.

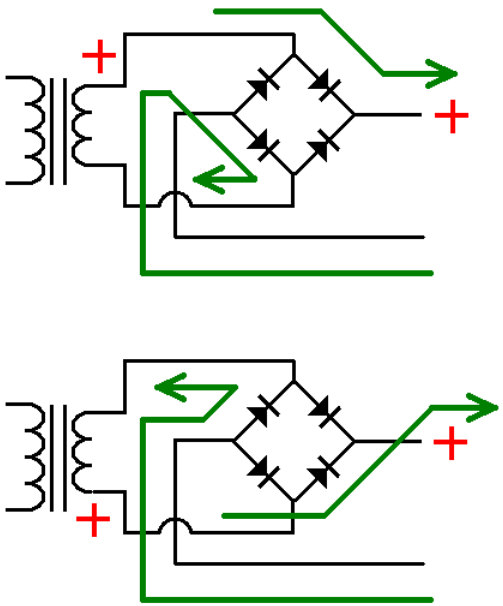
CHANGING AC TO DC

If we weren't connecting a filter capacitor, the output from the rectifier would look like half an AC sinewave, and in fact this is called **half wave** rectification. The output would get a lot smoother if we could make use of both positive and negative alternations. One way of doing this is to choose a transformer that has a secondary with twice the final voltage we'll need, and a center tap. If we connect the center tap to ground, we'll get **bipolar** AC from the two ends of the secondary – “bipolar” meaning that when one end of the secondary is positive, the other will be negative. This means that, at both positive and negative AC peaks, every half cycle, one end of the secondary or the other will be positive.



If we connect rectifiers between both ends of the secondary and the output, one rectifier or the other will always pass current through to the load, so both alternations will be present. We will no longer have that long stretch of zero voltage every cycle. This is called **full wave** rectification.

In many cases, we won't have a center-tapped transformer available. We can still get full wave rectification by using four rectifiers in a **bridge rectifier** configuration. During half the wave, two of the rectifiers conduct, connecting the transformer secondary across the output. During the other half wave, the other two rectifiers conduct, directing the positive and negative voltages to the same outputs as before. This is often called **diode steering**, as the rectifiers steer or guide the current so positive voltage always finds its way to the positive output. Take a pencil and follow the green arrows.



Wherever there's a rectifier in an electrical path, there's a voltage drop. Standard silicon rectifiers have a voltage drop of about 0.7V, or a bit more if you push them to their current limit. The bridge rectifier circuit has two rectifiers in the current path at all times, so the voltage drop is 1.4V. This isn't very important in, say, a 60VDC supply, but in a 5VDC supply the diode drop represents a serious loss. This can be substantially reduced by using **Schottky** rectifiers, which have a lot lower forward drop.

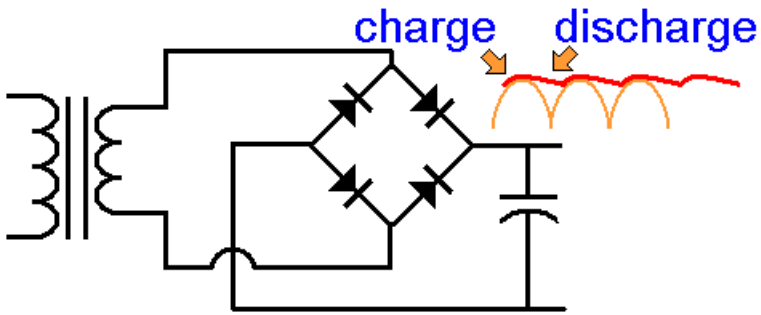
SMOOTHING THE DC

Most electronic circuitry requires smoother DC. There are several ways to turn the lumpy DC that comes from the rectifiers into a more constant voltage. This is called **filtering**, and when we're talking about a power supply, the word means both filtering in the usual sense that we studied in EM2, and energy storage.

CAPACITOR FILTERING

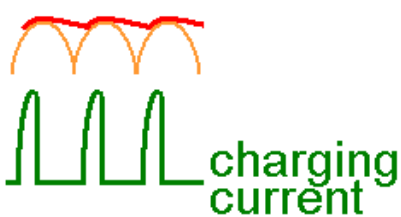
The simplest way of smoothing the lumpy DC is to attach a big capacitor across the output. “Big” may mean tens of thousands of microfarads. During the peaks of the waveform, current flows into the capacitor to charge it up.

The capacitor then fills in the gaps between the peaks, supplying current to the load and discharging slightly, until the next peak comes along and charges the capacitor up again. This is called **capacitor filtering**.



A city water tower works the same way: water is flowing out almost all the time, and every now and then a pump turns on and refills the tank. The tank is a reservoir, as is a power supply capacitor. It's called a **filter capacitor** because it filters out the AC part of what's coming from the rectifiers, leaving us with DC.

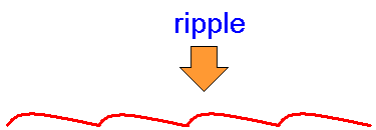
This isn't the most efficient process. Current coming through the rectifiers is zero except for a short period near each waveform peak, when it is very high. The only reason this usually doesn't cause a problem is that the power used by most electronic equipment is very small compared to things like refrigerators and stoves.



Capacitor filtering does have the advantage of being simple and inexpensive.

Note that the capacitor charges up to the peak voltage of the waveform, not the average or RMS voltage. This means the output of a capacitor-filtered power supply is higher than the RMS value of the AC voltage supplying it; specifically, around 1.4 times the RMS voltage coming from the transformer, minus diode voltage drops. On the other hand, the maximum output current capability is only about half the rated current output of the transformer.

The fact that the capacitor partially discharges between waveform peaks means that there will be a little residual AC in the DC. This is called **ripple** and with a full-wave rectifier, it has a fundamental frequency of 120Hz because recharging happens twice during each 60Hz cycle.



The larger the capacitor, the less it discharges between peaks and the lower the ripple component of the DC. However, it does not pay to try to smooth out the ripple by using a huge capacitor. If it discharges very little between peaks, that means all the recharging must occur during a very short period at the peak of each wave, resulting in short pulses of very high current.

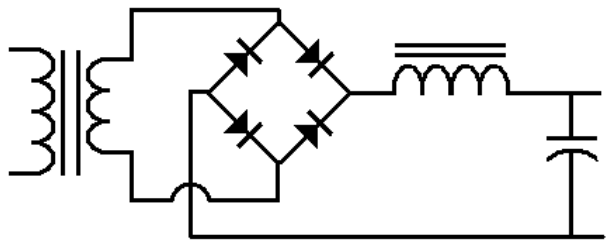
It's more practical to choose the capacitor to give reasonably smooth DC and then add other circuitry to smooth it further if necessary. At full load, the voltage should drop 5% to 10% between peaks. It's not difficult to figure out the required capacitor size, since we know from EM2 that change in voltage per second is current in amps divided by capacitance in Farads. Suppose we are building a 20V supply, the load will draw 2A, and we want the voltage to drop 5%, or 1V, between recharging peaks. 1/120 second is about 8mS, so:

$$1V / 8 \times 10^{-3} = 2 / C$$
$$C = 2 \times 8 \times 10^{-3} = 16 \times 10^{-3} = 16{,}000\mu fd$$

INDUCTOR FILTERING

Another way to smooth the DC is to put a large inductor in series with the output. Recall that an inductor opposes sudden changes in current flow the way a flywheel opposes sudden changes in RPM. An inductor used for power supply filtering is called a **choke** because it passes DC but chokes out the AC lumpiness.

Choke filtering is only practical with a fairly constant, permanently connected load. Changes in the load current will cause momentary changes in the DC voltage coming out of the choke. Suddenly disconnecting the load will cause a massive voltage spike in classic inductive flyback fashion. This problem can be eased, and the filter made to work with varying loads, by adding a filter capacitor after the choke. Even in this case, it is not a good idea to operate a choke-input filter with no load.



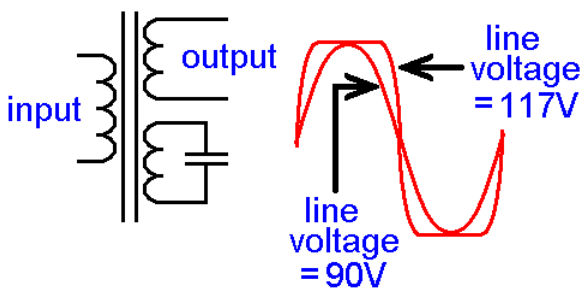
The output voltage of a loaded choke-input filter is equal to the average of the input waveform, minus diode drops – not the peak. The main advantage of a choke filter is that input current flows during all or a large part of the cycle. Its main disadvantages are that a choke for a 60Hz power supply is expensive and bulky and it radiates an alternating magnetic field that may cause hum in other parts of the circuit. This can be prevented by encasing the choke in a magnetic shield, but of course that makes it even more expensive and bulky. For these reasons, choke filters have fallen out of favor. However, they are still used where the high current pulses of a capacitor filter would be unacceptable. They've also made a comeback in switching power supplies, which we will study shortly. These operate at high frequencies, so the chokes can be small and cheap.

SWINGING CHOKES

I bet you just got a mental image of a choke wearing trendy clothes, hitting on that cute capacitor at the other end of the bar. Try to focus, please.

A swinging choke is used where the DC load varies. It's deliberately made to saturate at higher currents. As current increases and the choke begins to saturate, it acts less like a choke more like a wire. And that means, at higher current, the filter begins to resemble a capacitor-input filter, charging up to a higher than average voltage. This counteracts the tendency of all simple power supplies to **sag** – to lose voltage – as current increases. In other words, it's a crude voltage regulator. Crude is often good enough.

FERRORESONANT TRANSFORMERS



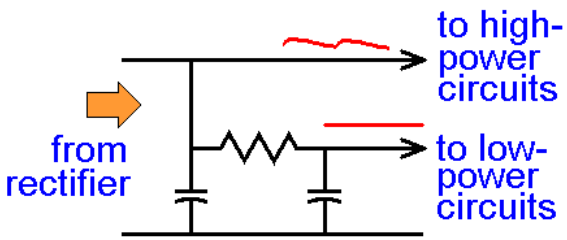
Another way of smoothing the DC is to start with a **resonant** transformer. This is a transformer with a separate winding on a separate part of a transformer. This winding is connected to a capacitor and forms a resonant circuit. Without going into a lot of detail, the transformer has a built-in regulating function so that, over an input range of perhaps 90V to 130V, and a wide range of loading conditions, the output is almost unchanging. Moreover, around 117V the output waveform resembles a square wave, which converts to DC much more efficiently. This is an old technology that is still used, even in new equipment, but not nearly as often as it once was. In power supplies that operate at high frequencies, resonant transformers are used for a different reason: to reduce radiation of interference.

MULTIPLE FILTERS AND DECOUPLING

An amplifier often has to deal with a signal level measured in microvolts. If the DC that powers this amplifier has half a volt of ripple on it, enough of that ripple will get into the signal to contaminate it beyond repair. So the DC from our basic power supply may have to be cleaned up further.

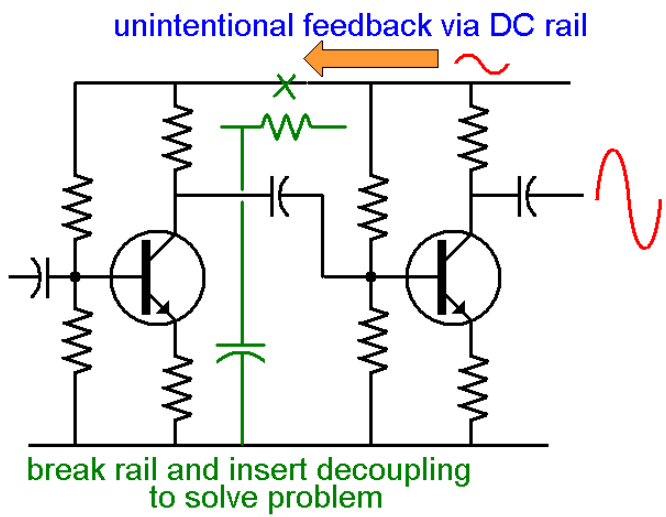
As a general rule, high-power parts of a circuit do not require as clean DC as low-power parts of the circuit. This is convenient, because the higher the power, the more complicated it is to filter and regulate. Where a piece of equipment has both low- and high-power circuits, the usual practice is to break the power supply rail in two (or more) pieces.

Unregulated, capacitor-filtered DC is fed directly to the high-power part of the circuit, while a regulating circuit further processes the raw DC to provide cleaner DC to the low-power part of the circuit. Thus regulation is only performed on a small part of the total DC.



A simple way to reduce ripple is to follow the capacitor with a resistor and another capacitor. This amounts to a lowpass filter. If we set the turnover frequency low enough – say, 0.1Hz – we'll get a lot of attenuation at the 120Hz ripple frequency. This method only works when the load current is measured in mA or less, because there will be a voltage drop across the resistor and it will dissipate power. Also, it doesn't work with a varying load, because then the voltage drop will vary and the output voltage will be unstable. In a great many cases these limitations pose no problem, because the parts of the circuit that require the cleanest power are those parts that draw small, unvarying amounts of current.

Splitting the power supply rail like this has another advantage. We've always treated power supply rails as if they're 100% “stiff,” but that's not really true. It is possible for the high current fluctuations at the output to cause voltage fluctuations on the power supply rail that couple backwards to the input circuitry. It doesn't take a lot of coupling to cause problems, because that's the most sensitive part of the circuit . The whole unit can wind up oscillating, usually at a low frequency. This is called **motorboating** because it sounds like an outboard boat motor at low speed. It's an aggravating problem to diagnose. There's no obvious malfunction anywhere in the signal path, because the source of the problem is not in the signal path, it's in the supply. Breaking the rail and filtering the low-power part separately prevents this from happening. This is called **decoupling** and a capacitor used in this way is called a **decoupling capacitor**.



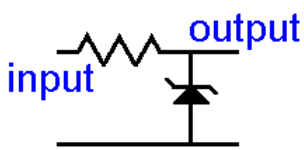
If you're called on to service an older piece of equipment that once worked well but now motorboats, it's likely that a power supply decoupling capacitor has dried out.

VOLTAGE REGULATION

Lots of electronic equipment doesn't require a precise DC voltage. As long as the DC is smooth and in the vicinity of the "correct" voltage, that's good enough. For circuitry that requires a tightly controlled voltage, a regulator is used.

The Shunt Regulator

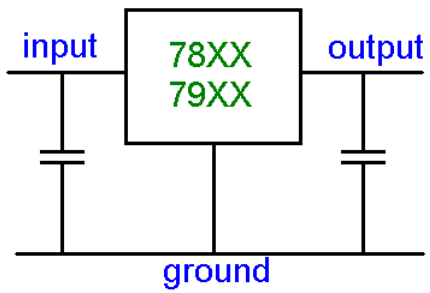
The most basic regulator is a combination series resistor and a semiconductor shunt device. This shunt passes as much current as required to drag the DC voltage down to the desired supply voltage. The classic shunt regulator is the **zener** diode, which we studied in chapter 2.



For a zener regulator to work, the current flowing through the series resistor must be greater than the maximum current that will ever be drawn by the load. This insures that there will always be some current flowing through the zener and it will always be able to perform its shunt function. Its filtering ability is improved by connecting a capacitor across the zener, typically a 100uf electrolytic. Shunt regulation is wasteful, but that hardly matters where only a few mA are involved. In a 100-watt transmitter, wasting 5mW at the input means increasing total power draw by 0.05% or less.

The Series Regulator

A more efficient regulation method, used at currents up to several amps, is to insert a series regulator between the source and the load. This regulator passes only as much current as the load actually draws, so there isn't a lot of wasted current as with the Zener. For many years, the most common type has been the "78XX" family of three-terminal regulators. "XX" refers to numbers that identify the voltage. For example, a 7805 puts out 5 volts, a 7812 puts out 12 volts and so on.



The three terminals are input, output and ground. The ground is necessary because the regulator measures its output against ground to maintain a constant output voltage. A negative feedback circuit is used, with the result that the output voltage is very stable.

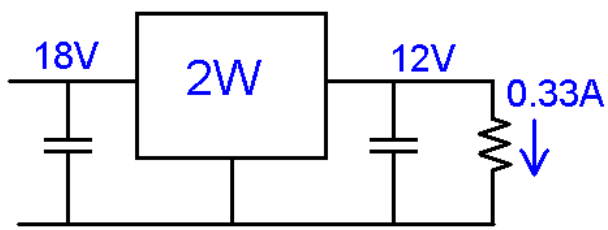
If you're building a negative-voltage power supply, there's an equivalent 79XX series of negative regulators. “78XX” and “79XX” numbers are not used by every manufacturer; some offer equivalent devices using their own numbering systems. There are also regulators whose output voltage can be adjusted with a potentiometer.

These regulators are integrated circuits, and take advantage of the IC's ability to offer additional features at low cost. They include temperature and current sensing circuits, and automatically shut down if the temperature or load current exceeds preset limits.

If you've read earlier tutorials on negative feedback, you know that when feedback is used, there is a chance that the circuit will go into oscillation. For 78/79 regulators to be stable, neither input voltage nor load can be permitted to change quickly. This means we have to add a little storage on either side of the regulator – capacitors between input and ground, and between output and ground. Any value between 0.1μfd and 1μfd is appropriate. They should be placed close to the regulator, so there is less than an inch of wire between regulator and capacitor. This is especially critical with the 79XX regulators, which are more prone to instability.

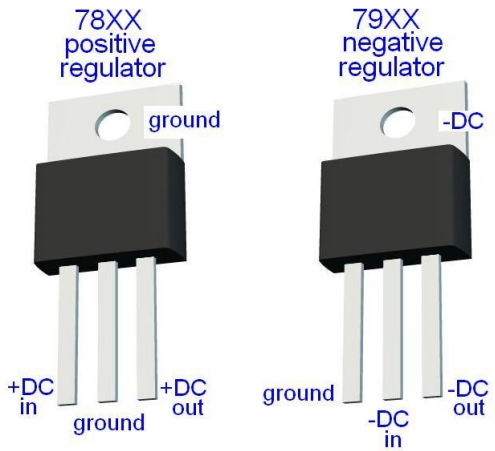
The regulator takes the DC from the basic capacitor-filtered supply and drops the voltage to whatever the regulated output is supposed to be. This means the DC voltage source voltage, including ripple, must always be higher than the regulator output, plus a safety margin that depends on exactly which series regulator you're using. More recent regulators can operate with lower safety margins, an important consideration with low-voltage, battery-powered circuitry. These are known as **LDO** or **low dropout** regulators.

The voltage-dropping operation also means that the regulator produces heat. After all, there's a voltage drop across it, from input to output, and current flowing through it, and that combination translates to heat. For example, if you have a 7812 putting out 12 volts, and the input is 18 volts, it's dropping 6 volts. If it's passing a third of an amp, it's dissipating 2 watts.



That's why most 78/79 devices have metal tabs. In any circuit where the regulator dissipates more than a watt, the metal tab is attached to the chassis or other heat sink.

Internally, the tab is connected to the most negative point in the regulator. For a 78XX positive regulator, that's ground, so the tab can usually be attached directly to the chassis. For a 79XX negative regulator, it's the DC input voltage, so the tab must be isolated from the chassis using insulating hardware and an insulating mica washer.

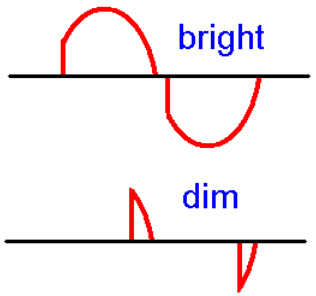


SWITCHING REGULATORS

I've said that there are two kinds of regulators, linear and switching. People have fallen into the habit of thinking of anything that involves switching as digital, but strictly speaking the switching regulator is an analog circuit.

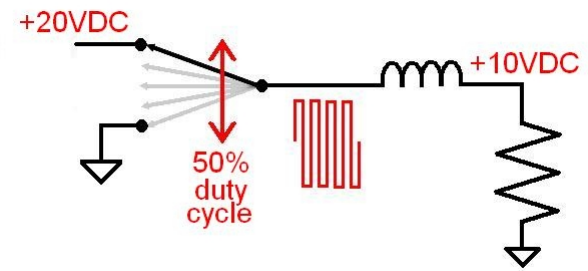
Let's start by recalling the discussion of incandescent lamp dimmers. In a triac dimmer, dimming is accomplished by varying the **duty cycle**; that is, by waiting until partway through each AC half cycle to turn on. Current does not flow during the first part of each half cycle, so the average current and thus the average power is reduced.

How much the average voltage is reduced depends on how long the circuit waits to turn the triac on during each half cycle.



A switching regulator is basically a lamp dimmer with refinements. First of all, if we want a DC output we have to start with the DC input from a standard capacitor-filtered power supply, the same as with an analog regulator.

In a lamp dimmer, current is switched on once each half cycle, and off when the sine wave goes through its zero crossing. In a switching regulator, switching occurs at a much higher frequency, at least 20kHz. The output circuit switches between the DC input voltage and ground like an SPDT switch. This is called **chopping** the DC voltage. In other words, the output is a rectangular wave with a duty cycle that is made to correspond to the ratio of input to output voltage.



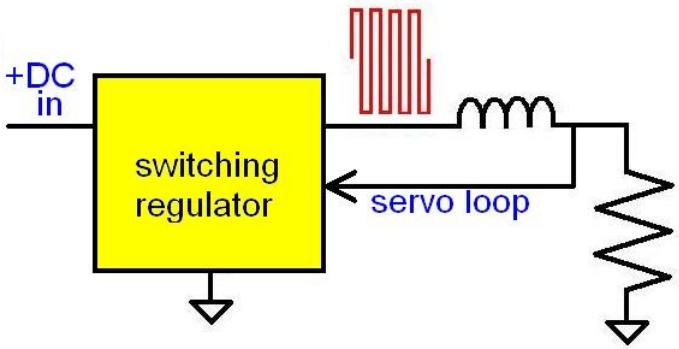
Since we don't have AC zero-crossings as we do with a dimmer, we can't use an SCR or triac as a switching device. We'll have to use BJT's, MOSFETs or IGBTs.

Electronic equipment requires smooth DC, so we have to add a filter. The simplest filter is a series inductor which takes the rectangular wave and turns it into a constant DC voltage with a little triangle-wave ripple on it. We add a shunt capacitor to prevent sudden changes in voltage and to further smooth the output, in the manner of an LC low-pass filter.

The reason for the high chopping frequency is that, the higher the frequency, the smaller the inductor and capacitor required to filter it into smooth DC. This has kept pushing chopping frequencies upward as faster transistors have become available. By today's standards, 20kHz is quite a low frequency. Some power supplies chop at 1MHz.

So now we have everything we need: a source of partially-filtered DC, a variable duty cycle, high-frequency chopper, and a smoothing filter. If you were awake when you were reading chapter 3 – not a certainty, given its length – you'll recognize that this is essentially a class D amplifier.

A switching power supply requires one thing that is not automatically necessary in a class D amp: a servo loop. We need to sample the DC output of the filter and feed it back into the switching circuit, so as to adjust the duty cycle as needed to get the average DC output voltage we want. There is no reason to design our own circuit to do this, as there are a great many switching regulator IC's available that also incorporate various protection and monitoring functions.



The final, filtered output of a switching regulator is seldom really smooth. It usually has very little ripple at 60 or 120Hz, due to that servo loop, but it often has a few mV of ripple at the switching frequency. With many kinds of equipment, including most digital circuitry, that doesn't matter. However, if better filtering is required, we can always follow the filter with another filter. In this application, choke-input filters are practical, because now that we're dealing with maybe 100kHz ripple instead of 120Hz ripple, the chokes can be small and cheap.

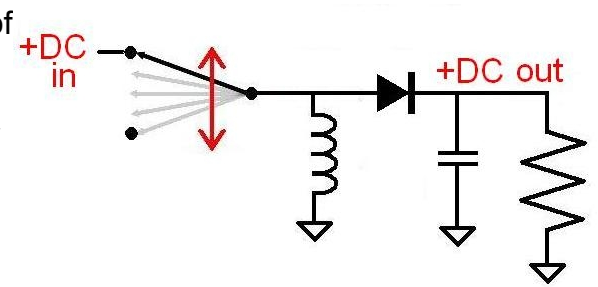
The great advantage of a switching regulator is that its power dissipation is very low – in theory, zero. The output is connected directly to either the +DC supply or to ground at all times, so there is no voltage drop in the switching circuitry and thus no dissipation. In the real world, there's no way to switch high power without some losses, but efficiency routinely reaches 90% or better. This also means there's almost no heat to get rid of. In battery-powered applications, the high efficiency means batteries last longer.

The main complication is that a 100kHz rectangle wave has great gobs of high-frequency harmonics in it, and these have to be suppressed to keep the regulator from radiating RF hash all over the neighborhood. This is especially critical in something like a cell phone, where crud from a regulator could swamp RF reception, defeating the purpose of the phone. Adequate filtering requires complicated circuitry, but at 100kHz, it's circuitry made of tiny, cheap components, so as long as the engineers know what they're doing, no problem.

Buck and Boost Regulators

The type of regulator described above is called a **buck** regulator because the inductor “bucks” or impedes the rapid voltage changes present at its input. The DC output voltage is equal to the voltage on the first filter capacitor or lower; it can never be higher. This will be obvious if you consider that the highest output voltage would represent a 100% duty cycle – the series switch would just turn on and stay on, connecting input to output. You can't get more “on” than 100% on.

There is a kind of regulator where the output voltage is higher than the input. This is called a **boost** regulator for obvious reasons. It involves an inductive flyback. For part of each cycle, a switch applies input DC voltage across an inductor, causing current to flow and magnetism to build up. Then the switch opens, causing the inductor to generate a flyback. If this flyback is captured by a diode and capacitor, we get a filtered DC voltage that can be as high as we want it. This is how a xenon flash unit generates the high voltages needed to operate. This is called a **boost converter**. If we add a servo loop that controls the duty cycle, we get a boost regulator, which adjusts the duty cycle to provide a regulated DC output voltage that is higher than the input voltage.



The longer during each cycle that current flows through the inductor, the more magnetism it will build up and the more energetic its flyback. (If you're familiar with automotive ignition systems, you'll recognize the similarity to “dwell time.”)

Battery-operated devices like cellphones use boost converters extensively, because LCD displays and piezo-type speakers require voltages much higher than the few volts provided by the battery. These circuits are also called **DC to DC inverters**.

Boost-Buck Regulators

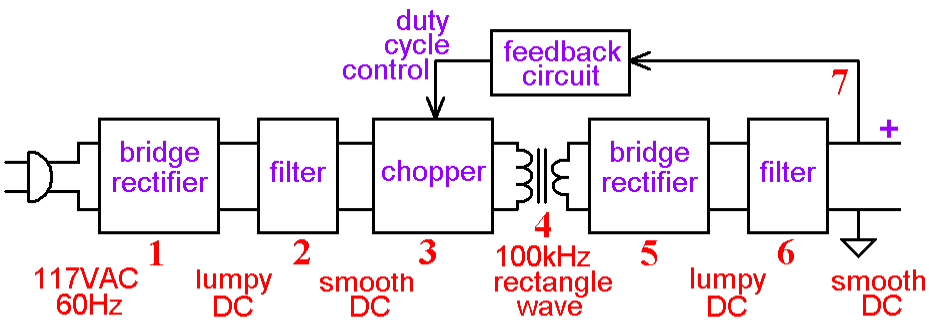
Boost and buck functions can be combined, so a regulator can boost when necessary and buck when necessary. This is a tricky process, especially around the transition voltage where the output is about the same as the input. But a lot of tricky design can be put into a cheap IC nowadays, and in any case it's not your problem! Just letting you know so that, if you ever have to work on a switching supply where the output DC is about the same as what's coming from the first filter cap, expect some complicated stuff.

Also FYI, there's a way of designing a boost-buck regulator that doesn't have the transition problem by starting with a negative DC voltage and turning it into a positive regulated DC output. Not many engineers do this, but in case you run into one of those, now you won't be confused.

And, just to cover all the bases: I've been talking about positive switching regulators, but of course the same functions can be done with negative voltages.

SWITCHING POWER SUPPLIES

A switching power supply works like a switching regulator, but it moves the switching function upstream.



1. Incoming 117VAC power is rectified to DC.
2. The DC is capacitor-filtered.
3. The DC is chopped to turn it back to AC at perhaps 100kHz.
4. A transformer steps the voltage up or down.

5. The new voltage is rectified to DC. Since we're dealing with a high frequency at this point, and often with a low voltage, the use of Schottky diodes is pretty much required.
6. A final filter removes 100kHz ripple.
7. A servo loop samples the DC output voltage and adjusts the duty cycle of the chopper as needed to obtain the desired output voltage. Note that the servo loop must include an isolation function to keep the DC output circuit separate from the AC line input.

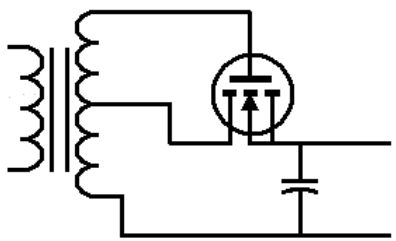
At 100kHz, iron laminations are not necessary or desirable (losses are too high). A switching power supply's transformer is almost always made with a ferrite core. The advantage of chopping the DC before the transformer is that a 100kHz transformer is a tiny fraction of the size, weight and cost of a 60Hz one. That's worth the extra complication, because a 60Hz transformer accounts for most of the size, weight and cost of a linear power supply.

Most modern switching supplies are based on integrated circuits that combine switching on the primary side of the transformer, duty cycle control, overcurrent protection and an impressive number of other functions.

SYNCHRONOUS RECTIFICATION

A lot of modern digital equipment involves very low voltages, 3 volts or lower, at high current. In this situation, the voltage drop of rectifiers, even Schottkys, results in unacceptable power losses.

The voltage drop can be all but eliminated by using power MOSFETs where we would have used rectifiers. MOSFETs are now available with “on” resistance of mere thousandths of an Ohm, which is about the same resistance as a couple of feet of 12-gauge wire. If we turn the MOSFETs on and off in step with the voltage transitions, we get the equivalent of a rectifier circuit with almost no voltage drop. This is called **synchronous rectification**.



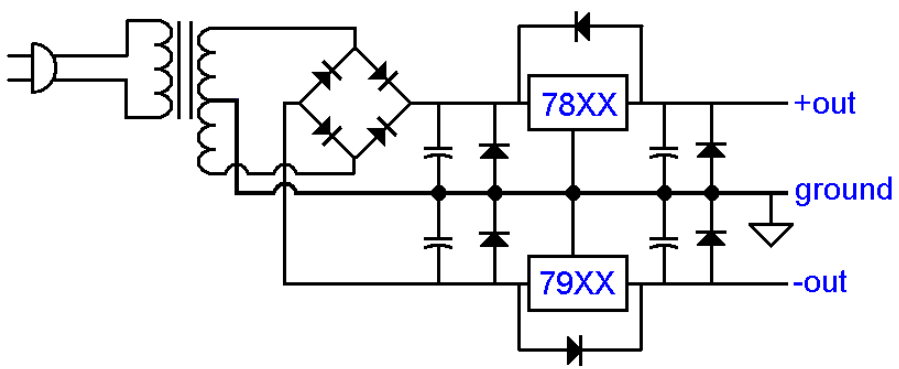
The graphic at right is a very simplified illustration of how this is done. Whenever the voltage at the transformer center tap is high, an even higher voltage turns on the FET. When it's low, an even lower voltage turns the FET off. Thus the FET works like a switch that's synchronized with the waveform, or a diode but without the forward drop.

Switching supplies that use it can achieve efficiencies of more than 90%, which is pretty impressive when you think of all that goes on in a switching supply and all the places where power could be lost.

MULTIPLE-OUTPUT SUPPLIES

There's no reason a power supply has to be limited to one output. Way back in the 1930s, radios typically had transformers with three secondaries – two low-voltage ones to drive tube filaments and heaters and one to provide for the high-voltage DC required by tubes.

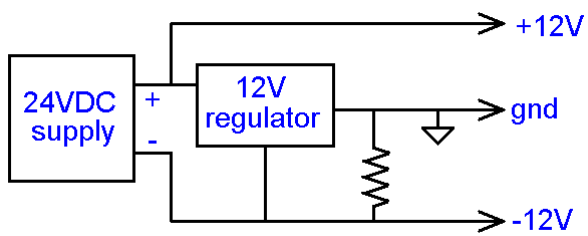
One kind of multiple-output supply is the **bipolar** supply which has one output that is positive with respect to ground, and another that supplies an equal negative voltage. This is a very common arrangement with analog equipment.



The rectifiers in the diagram are required to insure that no regulator terminal is pulled past ground in either direction, even during startup. If that were allowed to happen, the regulator would probably be destroyed.

Computers usually use switching supplies that provide + and – 12VDC, +5VDC and very often +3.3VDC. The actual number-crunching circuitry uses the +5V or +3V power. This accounts for most of the power, several hundred watts of it. The +12V supply is there to drive fans, and the -12V supply to drive some of the interface circuitry.

It is possible to create a bipolar supply via a **derived ground**. A voltage regulator is used to provide a voltage approximately halfway between the positive and negative supply rails.



This voltage, not the negative rail, is connected to the chassis and circuit grounds and becomes the reference for the circuitry. It's a little tricky, you have to watch current flows etc. It's something only an experienced engineer should design, but I wanted you to be able to recognize a derived ground when you run across one.

POWER MANAGEMENT ICs

Sophisticated power-control ICs have become a big business; the Mouser catalog lists around 65 pages of them. In AC line operated, high-power gear, power management can extend equipment life and save money. In battery operated, low-power gear, power management can extend a battery's charge. Do you own a laptop or tablet? How long does a battery charge last? However long that may be, it's longer than it would have been without power management chips that shut down parts of the circuit not currently in use.

A sampling:

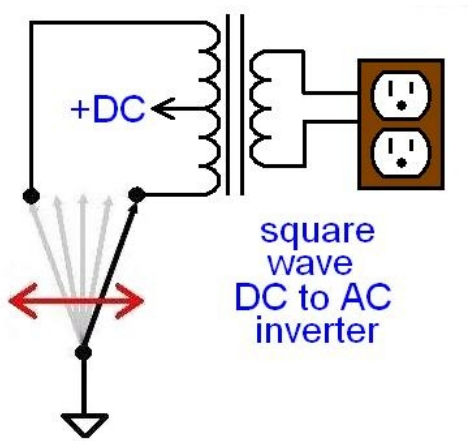
- Power sequencers: automatically turn several units on in a sequence, and off in a reverse sequence.
- Hot swap controllers: **hot swapping** means plugging and unplugging parts of a system while power is on. This can lead to damage including blowing up the power supply that runs the whole operation. Hot swap controllers limit the current and sequence the switching to prevent damage.
- References: precision voltage sources.
- Ground fault interrupt: the chips behind the buttons in a GFI outlet, monitoring incoming and outgoing current and tripping a breaker if they're not the same.
- Power factor control: compensating for loads that lead or lag the phase of the AC power line. We'll be talking a lot about this in EM5.
- Standby power management
- Battery management: modern batteries like Lithium-Ion are wonderful, but they require very sophisticated charge and discharge monitoring.
- Charge pumps: one way to obtain a DC voltage higher than the power supply source voltage, or of the opposite polarity, discussed below.
- USB and power-over-ethernet controllers: it's handy to be able to draw small amounts of power from these computer connections, but an overload could shut down the computer. These chips manage and limit the power draw from ports.

INVERTERS

Except in the case of a DC-DC inverter, the word **inverter** usually refers to a unit that produces an AC output. It's a backwards power supply which takes DC battery voltage and turns it into 117V, 60Hz AC for the purpose of powering 117VAC equipment from a battery.

In its simplest form, an inverter is just a step-up transformer and an SPDT switch that chops the DC into AC. This physical-switch arrangement was actually a very common technology at one time.

The switch contacts were mounted on a piece of metal spring material that was set to oscillating by an electromagnet. Every car radio used it from the 1930s until transistors replaced tubes in the 1960s.

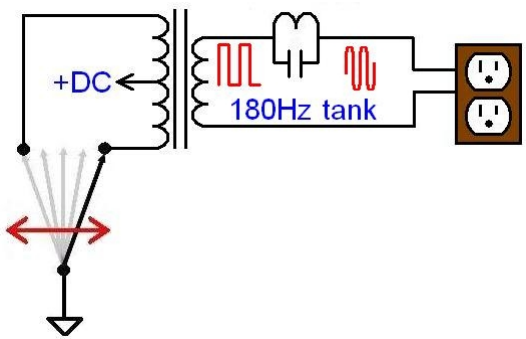


Square-Wave Inverters

In the modern inverter, the switch contacts are replaced with transistors which are driven by opposite-polarity square waves. It is important for the square waves to have exactly equal duty cycle, or there will be a net DC current through the transformer, which will tend to saturate it.

The problem with a square wave inverter is of course that its output is in fact a square wave, which a lot of electronic equipment doesn't like. It's fine for lamps and shavers and such, but motors and electronic power supplies expect to see a sine wave. Also, recall that a capacitor-filter DC supply charges up to the peak of the sine wave. When powered by a square-wave inverter, the DC voltage will be roughly 30% lower than with a sine wave, and the unit may not work.

Trapezoid-Wave Inverters



These are also called **pseudo-sine wave inverters**. There are several kinds of pseudo- sine inverters. In one type, an LC tank is inserted in the circuit. The tank is tuned to 3X the inverter frequency, or 180Hz. This tank takes up energy during part of each half cycle and gives it back during another part, causing the square wave to morph into something crudely resembling a sine wave.

True Sine-Wave Inverters

The reason square- and trapezoid-wave inverters exist is that it used to beexpensive to create high-power sine waves. Sine wave inverters required either multiple taps on the transformer primary and complex switching circuitry, or a large, expensive LC filter tuned to 60Hz. The main difficulty was not getting this to work but getting it to work with a wide range of different kinds of loads, and keeping things from momentarily going nuts when a load was turned on or off.

Now that we have transistors that can be turned on and off at rates of hundreds of kHz, a sine wave can be produced by creating a high-frequency rectangle wave and varying the duty cycle. A switching sinewave inverter is just a class D amplifier with a 60Hz sinewave connected to its input. This now permits manufacturers to produce true sine wave inverters at reasonable prices.

This technology can be applied to lots of other things:

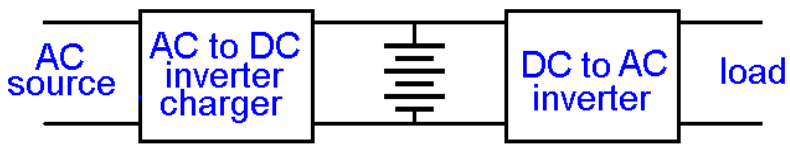
1. Lamp dimmers that put out smooth AC, instead of the crude chopped waveform of traditional SCR and triac dimmers.
2. High power audio amplifiers that are light and compact and generate very little heat.
3. AM transmitter modulators that are much smaller, run much cooler, and cost less to operate than those of the past.
4. DC and AC power supplies whose output voltage and frequency can be digitally controlled.
5. ...etc.

UNINTERRUPTIBLE AC POWER SUPPLIES

There are many places where a power outage would be a very bad thing – large data centers, emergency communication centers, hospitals, police stations. In these situations, the most critical electronic equipment is powered by a **UPS** or **uninterruptable power supply**. A UPS is a battery and an inverter that switches into action when the building AC power fails. There's also a charging circuit to keep the batteries at full charge between outages.

UPS's are divided into two categories. In the first category, the inverter circuitry does not kick in until a power outage actually occurs. There is a delay of a fraction of a second, during which power to the load equipment goes down.

In the second category, the inverter is always running, supplying power to the load equipment. The charging circuit puts enough power into the batteries at all times to make up the power drawn out by the inverter.



This is a more expensive approach and involves some wasted power due to the inefficiencies of the AC→DC→AC conversion process. However, with modern technology, the wasted power is a small fraction of the total. In any case, there is often no choice: in a hospital, for example, the power cannot be permitted to fail even for a moment.

A UPS also performs another important function: it filters and cleans up the AC coming from the power line, keeping radio-frequency crud and voltage spikes on the AC line from reaching the load equipment.

Large machinery often requires fail-proof power for safety reasons. For example, the power system for a newer elevator is essentially a big UPS. AC power is used to charge a whole closet full of batteries, which in turn power the motors. Since this power system is already in place for safety reasons, there's no reason not to run the elevator off the batteries and inverter all the time. We've already got the rectifier and batteries, so why not? This permits better control of speed and torque, which makes for a faster and smoother ride.

Often, the DC to AC inverter is omitted, and DC motors are used which are powered directly by the batteries. In some applications, DC motors simply work better.

VARIABLE FREQUENCY AC MOTOR DRIVES

One of the more aggravating problems associated with AC motors, which we'll study in more detail in EM5, is that they tend to spin at rates governed by the 60Hz power line frequency. It's much harder to vary the speed of an AC motor than a DC motor.

MOSFET transistors capable of switching very high power have recently become available. Manufacturers can now build **VFD's** or **variable frequency drives** which can turn 60Hz power into any frequency at any voltage. This finally permits something people have wanted for ages: a convenient way of varying AC motor speed.

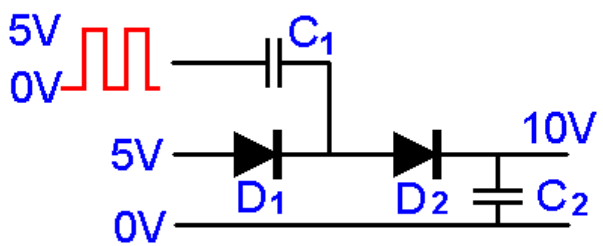
The concept is simple: a VFD is an AC-to-DC power supply, followed by a DC-to-AC inverter. There's no law saying the DC-to-AC converter must put out 60Hz, the output can be any frequency we care to generate it. The circuit is a class D amplifier with a variable frequency sinewave signal source. Today, the problems involved with high-power, high-frequency switching are well under control and VFD's that can drive multiple-horsepower motors cost only a few hundred dollars.

DC-DC INVERTERS

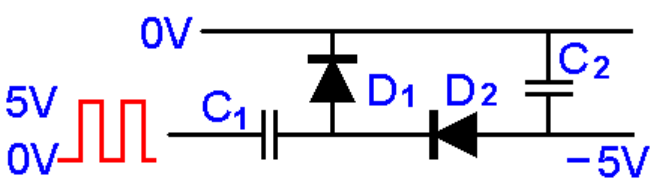
It's often necessary to derive a DC voltage that is higher than the system +DC rail. You've already seen one example, the boost regulator. Here are several others which do not provide a regulated output voltage, though with a little extra circuit complexity they can be made to. In every case – rule of thumb – turning a low DC voltage into a higher one involves chopping it into AC, then rectifying the result.

THE VOLTAGE DOUBLER

In the circuit below, I've ignored voltage drops across the diodes. The circuit includes a square wave generator, not shown, which switches between the 5V and ground rails. When the output of the square wave generator is low, current flows into C_1 through D_1, charging it to 5V. When the square wave output goes high, it lifts the negative end of C_1 to 5V, which means the positive end of C_1 will go to 10V. At this point current will flow from C_1 through D_2 into C_2, charging it to 10V. Thus the supply voltage has been doubled. The circuit is also called a **charge pump**.

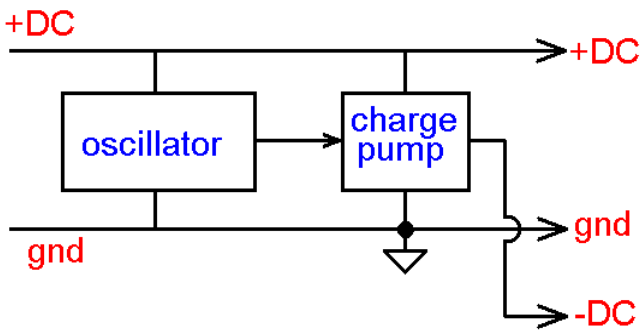


We can also turn a single-ended supply into a bipolar one by rearranging the components, as shown below.



We can continue to raise the voltage by adding capacitors and diodes and pumping up the charge as high as we want. This used to be a popular way to generate the thousands of volts required by an oscilloscope CRT.

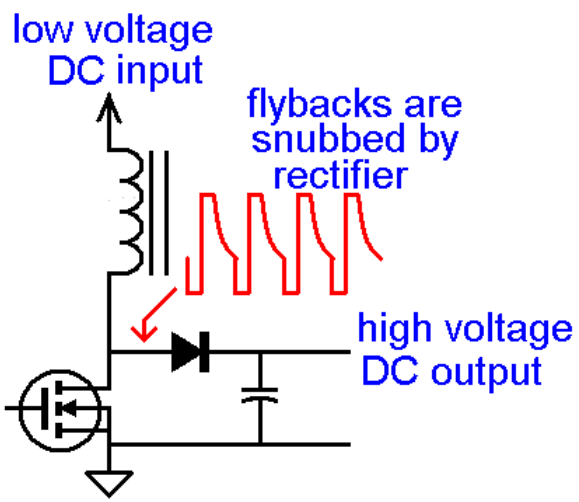
However, at low supply voltages, the diode voltage drops result in serious loss of efficiency. For this reason, charge pumps are seldom used to do more than double a voltage, unless an extremely high voltage is required for, to name one example, lightning research. Having seen a 40-foot electrical discharge *inside* a building (and not wet myself, thank you very much, though it was close) I can attest to the effectiveness of diode voltage multipliers.



Probably the most frequent use of charge pumps today is to provide a voltage that a circuit needs but the power source doesn't supply. For example, most digital circuitry runs on a single positive voltage. If a particular IC needs a negative voltage, it may very well have an oscillator and charge pump circuit built right into the chip.

INDUCTIVE FLYBACK BOOSTER

You've already seen something a lot like this; the transistor is turned on long enough to get some current flow going in the inductor, then turned off to generate a flyback. Widely used where a DC voltage of 50-300V is needed and precision isn't very important.



When we get to EM5, you'll see this circuit again. Being able to control the duty cycle permits us to build a DC supply that draws power smoothly from the AC grid, instead of in pulses. That's really important when it comes to making the best use of our power grid, which is reaching its limits. And if that isn't an obvious plug to get you to buy the next study guide, I don't know what is.

End of chapter 9

EM3 Chapter 10
OPTOELECTRONICS

Objectives of this chapter:

- *define "optics" and "optoelectronics"*
- *describe the three classifications of light and corresponding wavelengths*
- *discuss the difference between photoconductive and photovoltaic devices*
- *discuss photoresistive devices, their advantages and limitations*
- *explain photoconductive and photovoltaic uses of a photodiode, and the advantages and limitations of each approach*
- *discuss bulk photovoltaic devices*
- *discuss light emitting diodes, and the voltages that different-color LEDs typically require*
- *explain how white LEDs operate*
- *explain the difference in life expectancy between white LEDs and other types*
- *discuss lasers and their uses*
- *define and discuss "optocoupler," "opto-isolator," "opto-interrupter"*
- *discuss fiber optics, the materials used, and what is meant by "window"*
- *define and discuss light amplifiers*

Optics is the science of light. **Optoelectronic** means an electronic device that creates or senses light. Light is like radio frequency energy, but the frequencies are much higher. The highest radio frequency most people deal with is around 2GHz or 2×10^9 Hz, used by cellphones.

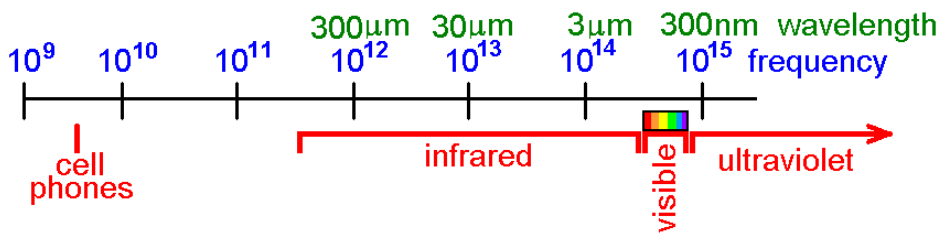
RF begins to become infrared light around $^1/_3$ terahertz or 3×10^{11}Hz. The visible light spectrum spans from red at about 4×10^{14} (200,000 times higher than the 2GHz cellphone band) to purple at 8×10^{14}Hz. X-rays are higher in frequency and gamma rays still higher.

Because these numbers are so huge, and because a lot of the science of light revolves around the physical size of light waves, we tend to measure light according to its wavelength rather than its frequency. As you know, the higher the frequency, the shorter the wavelength. Wavelength is speed divided by frequency:

$$\lambda = V / F$$

where wavelength is denoted by the Greek letter lambda, λ, and frequency is measured in Hertz. Velocity, in the case of light and radio waves in air or space, is usually measured in meters because it makes the calculations easy: the speed of light in space is almost exactly 3×10^8 meters/second.

Since we're using the metric system to measure velocity, we'll also measure wavelength in meters. People who deal with optoelectronics are in the habit of working in **micrometer** and **nanometer** units. A micrometer is 10^{-6} meter; a nanometer is 10^{-9} meter. Note that the abbreviation for meter, **m**, is lowercase.



You recall that sound is divided into three frequency ranges: frequencies we can hear called audio or sonic frequencies, frequencies below what we can hear called infrasonic frequencies, and frequencies above what we can hear called ultrasonic frequencies.

Light wavelengths are similarly divided into:

- Visible light: about 800nm (red) to 400nm (violet)
- Infrared light: longer wavelengths than visible red light, 1,000µm to 800nm
- Ultraviolet light: shorter wavelengths than visible violet light, 400nm down to 10nm

The actual light spectrum, of course, is continuous; we just make these divisions because they're useful to us.

Infrared light is further divided into:

- Near infrared, meaning near the visible spectrum: 800nm to 3µm
- Mid infrared, lower frequency/longer wavelength: 3µm to 50µm
- Far infrared, far below the visible spectrum: 50µm to 1mm

Ultraviolet is divided into:

- Near ultraviolet, 400-300nm
- Middle ultraviolet, 300-200nm
- Far ultraviolet, 200nm and shorter.

The designations for ultraviolet wavelengths are not hard and fast; in many cases, ultraviolet is just divided into “near” and “far,” and the dividing point depends on what's most useful in a particular situation.

PHOTOSENSITIVE DEVICES

We'll start with devices that sense light. These are divided into two categories: those which control electricity, like valves, and those which make electricity, like generators. Some devices can be used either way. All electronic light sensors fall under the general heading of **photodetectors**.

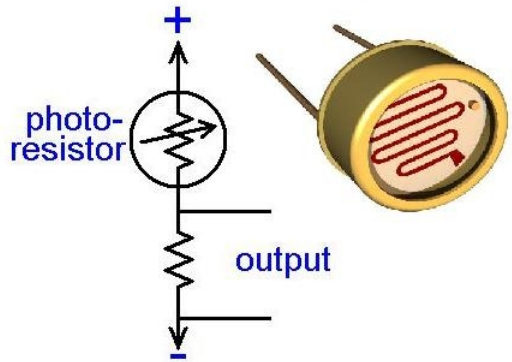
Valve-like sensors require an external power source to work. You recall that an ohmmeter measures resistance by placing a voltage across a circuit and measuring the resulting current flow. Similarly, to use a valve-like sensor, we place a voltage across it and measure current flow. In the dark, current is very low; in the presence of light, the sensor begins to turn “on” and allow current to flow. The brighter the light, the greater the current.

Generator-like sensors turn light directly into electricity. Some types generate just enough to feed into the input of a sensitive amplifier. Other types cover rooftops and generate kilowatts.

A device that operates like a valve is called a **photoconductive** device. A device that works like a generator is called a **photovoltaic** device.

Photoresitors

A photoresistor or **LDR** (**light-dependent resistor**) is a material, usually **cadmium sulfide** or **CdS**, that has high resistance in the dark and much lower resistance in light. The wiggly shape of the active element is due to the need to get a fairly long string of photoresistive stuff in a small case. Photoresistors are rather slow to respond, too slow to serve as a detector for an optical movie sound track for example. They do have one advantage: they are true resistors, and their voltage/current ratio is almost as linear as a regular resistor. For this reason, they are often used in AC voltage divider circuits to vary the amplitude of a signal where waveform distortion is not acceptable. BJT and FET variable-gain circuits with low distortion can be made, but it's much more complicated.

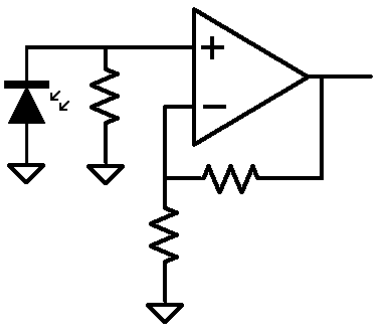


Photodiodes

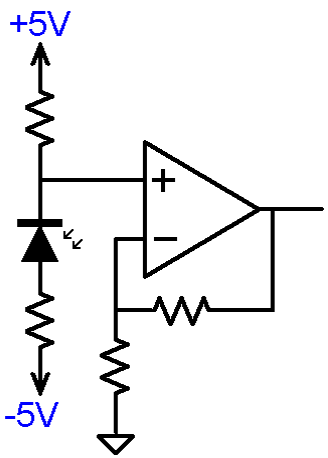
Strictly speaking, *every* semiconductor junction is sensitive to light. In most cases this is an undesirable characteristic, which is one reason so many solid-state devices are encapsulated in black plastic. A photodiode is simply a diode in a transparent casing whose light sensitivity is optimized and calibrated. One type, the **PIN** photodiode, so called because it has a thin extra layer of intrinsic silicon between the P and N regions, is used where high speed is important. A PIN diode's response time is typically around 20 nanoseconds, and can be made faster by circuit tricks.

In the symbol, the two little arrows represent incoming light, indicate a photodiode.





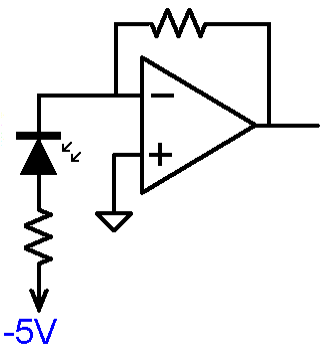
A photodiode can work in either photovoltaic or photoconductive mode. When in photovoltaic mode, the diode is simply connected to an amplifier in the normal fashion. It is most sensitive in this mode, its light input vs. voltage output characteristic is linear, and it has low noise. The drawback is that it is comparatively slow. Remember from our discussion of varactors that a diode's capacitance is highest when voltage across it is lowest. High capacitance means low speed. Thus a photodiode is used in photovoltaic mode where high speed isn't critical. (Circuits shown are simplified; real photodiode amps are more complex.)



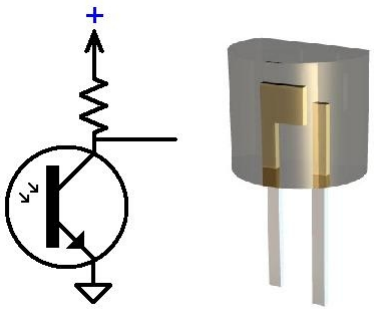
When high speed is important, a reverse bias is applied across the diode and it's used in photoconductive mode. The reverse voltage expands the depletion region, which is like pulling the plates of a capacitor further apart; capacitance is reduced. Now that there's voltage across the diode, we will have a small leakage current to contend with.

Photodiode current levels are very small, on the order of microamps. It's tricky to amplify such a small signal; moreover, at such a low current levels, diode and circuit capacitances slow things down. Photodiodes are high-speed devices, but it takes some ingenuity to retain high-speed performance *in a circuit*.

This problem can be eased by using an amplifier with a virtual-ground input. Because the load impedance across the diode is very low, the shunt loading of the diode's capacitance is insignificant by comparison. A photodiode system can be designed to work to 100MHz without much trouble, and heroic engineering efforts can get them to work into the gigahertz region.



Phototransisotrs



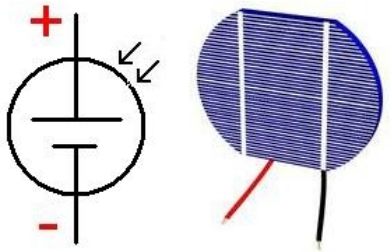
A phototransistor combines the functions of a conductive photodiode and a transistor amplifier. The main virtue is that the output current is much stronger, mA instead of μA, and in the case of a **photodarlington** transistor, current may be in the neighborhood of an amp. The drawback is that phototransistors are slow. That's of no concern when using a phototransistor to operate a door opener, but it becomes a problem when handling high-frequency analog or data signal streams. Many phototransistors can't even do a good job at higher audio frequencies, let alone RF. They can be sped up a lot by connecting a resistor between base and emitter, but this greatly reduces sensitivity. In many cases, the base is not brought out to a lead you can connect to.

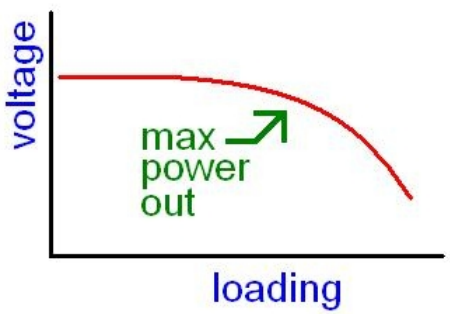
Opto-FETs and Photo-SCRs

These are simply optically activated FETs and SCRs. In the presence of light, they turn on in the same way regular FETs and SCRs do when a control signal is applied. For some reason, they aren't much used in discrete form, but they are often found in optocouplers, discussed below.

PHOTOVOLTAIC DEVICES

A **photovoltaic** or **PV** panel on a building, also called a **solar cell**, is a large silicon PN diode. In movie projectors, small PV cells are used to pick up the optical sound tracks.





A PV cell is like any other power source in that there's a limit to the practical load that can be connected to it. Beyond a certain point, additional loading results in less power output, not more.

Direct sunlight at sea level contains about 1kW of power per square meter, of which a little over half is infrared, a little under half is visible light, and about 3% is ultraviolet.

At the moment, the best solar panels capture about a third of that power. By the time they're assembled into arrays and interconnected, conversion efficiency is about 20%. Note that the symbol resembles a battery in a bottle, with the long line positive.

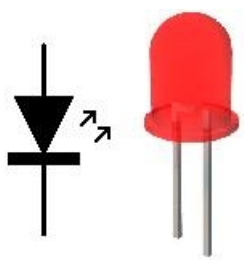
I am not going to discuss solar cells in detail because so much research is being done, and the science is changing so fast, that anything I say will be obsolete by the time you read it. I suggest downloading the ETA solar certification competencies; there's a list of up-to-date reading material included. We will be dealing a bit with solar panels in EM5, when we talk about power generation and distribution.

PHOTOEMISSIVE DEVICES

"Photoemissive" means a device that emits – creates – light. By 2020, incandescent lamps and even fluorescents will probably be museum artifacts. Solid state light sources will be everywhere.

Light Emitting Diodes

LED means **light-emitting diode**. Forward-bias it, so current flows through, and it will light up. The first LEDs were red, and they weren't very efficient. But during the 40 years since, LEDs have been constantly improved until today, you can get high-powered LEDs in any visible color as well as infrared and ultraviolet. In the schematic symbol the two little arrows, pointing outward this time, represent the light emitted from the LED.



Like any diode, a LED doesn't conduct at all until a certain threshold voltage is reached; beyond this point, a small increase in voltage results in a large increase in current. In other words, like any diode, it's hard to set the current by controlling the voltage. It's poor practice to connect a LED directly across a battery – especially a red LED, which has very little "give" in its V/I curve – because then there's no mechanism for controlling current except luck. Manufacturers of cheap imported flashlights do it all the time, but then, those flashlights have a habit of burning out.

The simplest way to control current is to put a resistor in series with the LED. LEDs have fairly predictable dropping voltages which depend on color.

- Infrared, red, orange and yellow LEDs are about 2-volt devices
- green LEDs operate on 2-3 volts
- blue, violet and white LEDs operate on 3-4 volts.



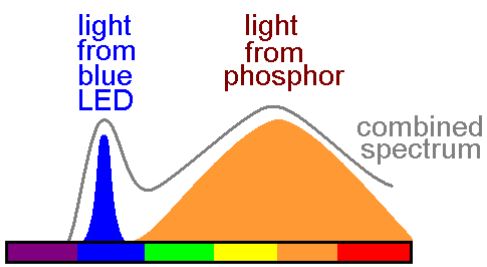
To find the resistor value, we first subtract a LED's operating voltage from the power supply voltage to get the voltage drop, then using Ohm's Law we calculate the resistance. Say we want to put 10mA thru a red LED running on a 6V battery; that's 4V / 10^{-2}A = 400Ω.

Of course, a dropping resistor implies wasted power. There are many efficient solid state control circuits available which use class D switching to control current.

Standard LEDs can handle up to about 30mA. If this limit is greatly exceeded for any length of time, they will melt and stop working.

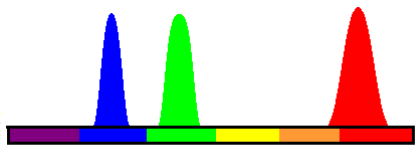
In the last few years, much higher-current LEDs have become available. As this is written (2012) the brightest LEDs will handle several amps, and a power level of 10 watts. These are so bright it's dangerous to look at them.

Most LEDs have an essentially unlimited life expectancy if they aren't driven beyond their specified current limit. The exception is white LEDs, which are blue LEDs coated with a warm-white phosphor. These LEDs have a limited lifespan because the phosphor eventually becomes exhausted. Still, white LEDs driven continuously at full rated current will probably take at least 5 years to fall to 50% of their original brightness.



spectrum of white LED

There's another way of obtaining white light from LEDs: group red, green and blue LEDs together. To the human eye, a mix of red, green and blue light looks like white light. Some manufacturers make 3-in-1 LEDs, known as **RGB** LEDs. However, where high-quality color rendering is required, 3-color LEDs do a poor job of revealing the full range of colors of an object. That's one reason why they haven't found favor in merchandise displays and similar applications.


spectrum of RGB LED

The best LEDs are about 30% efficient, which is at least five times the efficiency of an incandescent lamp and a bit better than a fluorescent lamp. They do not have the hazardous-waste disposal problem of fluorescent lamps.

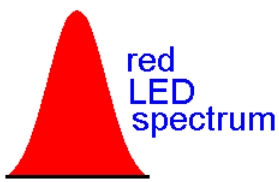
LEDs can transmit data by simply modulating the current though them, which varies their brightness. Phosphor-based white LEDs can only be modulated to a few hundred Hz, but the others will operate at much higher frequencies.

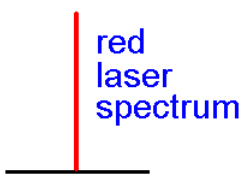
LASER DIODES

Most light sources create light that spans a certain chunk of the frequency spectrum. White light contains energy at every frequency in the visible light spectrum; it isn't too much of a stretch to compare it to noise in that sense.

Most colored light sources also emit light across a range of frequencies, the range is just narrower. There are a few light sources whose spectra are very narrow; for example, sodium lamps radiate in a narrow orange-yellow band. That's why colors look washed-out under sodium lamps; there just isn't any energy at other wavelengths.

But a laser emits light in such a narrow spectrum that it's usually considered a single wavelength. This is called **coherent** light. (In many kinds of lasers, the waves are also perfectly aligned in a beam, parallel to each other. That's called **collimated** light. The two terms are often confused.) Since there's only one wavelength, the waves stay in step as they travel away from the laser.



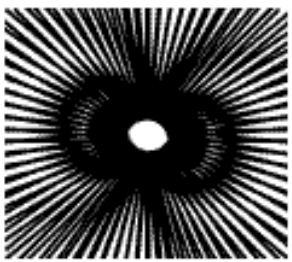


This can lead to some odd results. If you've ever looked at the spot made by a laser pointer, you have probably noticed that it's grainy, and the grain pattern shifts as you shift your gaze. That's not grain in the laser.

It's grain in your eyes, the grainy structure of your retinas, reacting with the laser light.

With some types of laser, the wavelength is determined by the materials, but with others the wavelength can be “tuned” during manufacture. This is fortunate because, as noted below, optical fibers only transmit light efficiently at certain wavelengths. Diode lasers transmit data the same way LEDs do, by varying the current through them.

You're probably familiar with Moire patterns, where small-featured images laid on top of each other produce large-featured interference patterns. There's a related technique used with lasers called **laser interferometry**. A laser beam is split in two, then recombined with itself using mirrors and prisms. At certain points the waves reinforce, and at others they cancel. (Waveform cancellation is called **destructive interference** and we'll study it in EM6.) This technique can be used to measure dimensions or distances with a very high degree of accuracy.



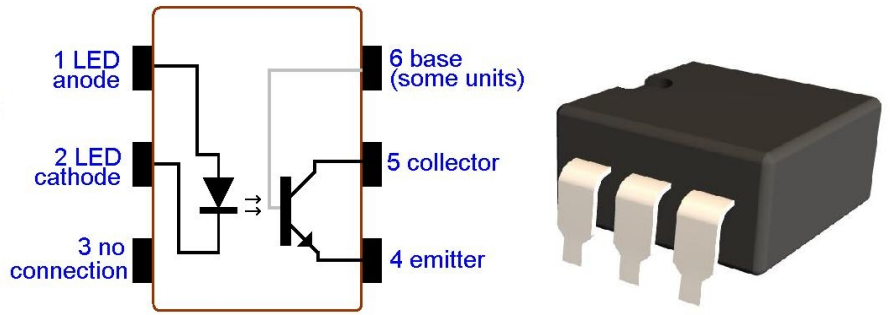
COUPLED OPTICAL SYSTEMS

As cool as solid-state light sources and detectors are, putting them together in pairs is even cooler. It gives us ways to get signals from one place to another in ways we couldn't before, and without problems we have with wire.

The main benefit is that we can get a signal from one place to another with no actual electrical connection; in other words, we can achieve isolation. A transformer does that too, but doing it with a LED and phototransistor has three important advantages: 1) a good optical coupling system is a lot cheaper than a good transformer; 2) an optical system will work with DC; 3) the transformer's electromagnetic radiation and sensitivity problems go bye-bye.

OPTOCOUPLERS / OPTO-ISOLATORS

An **optocoupler** combines a LED and an optical sensor, most commonly a phototransistor, in one package. If the right circuitry is connected, a current through the LED will cause a current through the phototransistor. This gives us a way to get a signal from point A to point B without any actual electrical connection. Very handy when point A is, for example, a computer and point B is a high-voltage, high-power machine. Because of this isolation function, optocouplers are also called **opto-isolators**. A pretty good optocoupler costs about 30 cents in small quantities.



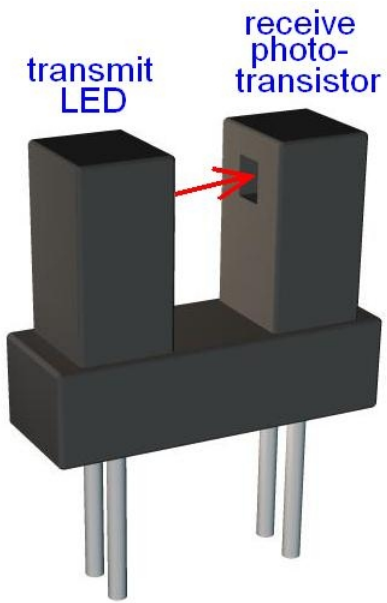
Optocouplers aren't terribly linear; they cannot pass a signal without distorting it a bit. They're also noisy. For this reason, they aren't much used in applications where the signal waveform must be preserved intact, such as analog audio. On the other hand, with control signals, RF and digital data, we don't care much about the precise shape of the waveform. All we care about is that we can recover that data that's encoded on it. Those are the kinds of applications where optocouplers are at their best.

Optocouplers are available with all the usual detector types – photodiodes, phototransistors, photodarlingtons, photo-SCRs and triacs, photo FETs. They are available in dual and quad versions, like opamps.

Opto-Interrupters

An opto-interrupter is an optocoupler that's made with an opening between the LED and the detector, so objects can be passed between them to interrupt the light path. This gives us a convenient way to, for example, count gear teeth per second and thus the speed at which a gear is turning.

In machinery, an opto-interrupter can often be used in place of a mechanical switch; it will never wear out, get stuck or go out of adjustment.



If we want to stretch a point, we could call a widely separated light source and detector an opto interrupter – a burglar alarm across a doorway, for example. In theory, there's no limit to how far the source and detector can be separated, but of course, in reality the light disperses and gets too weak for the detector to pick up, or gets swamped by ambient light.
We can improve things by modulating the light with an on-off pattern, then building the detector to look for that pattern. That's how a TV remote control with a single infrared LED is able to work in a brightly lit room.

FIBER OPTICS

A fiber-optic system can be viewed as an optocoupler with a large distance between emitter and detector – many miles, often. Glass or plastic fibers contain the light and guide it from emitter to detector. Short lengths tend to have infrared LED sources, and the light conductor tends to be a bundle of plastic fibers. When sending light over long distances, single glass fibers and lasers are used. Fiber optics are a specialty within the larger field of electronics, and many techs, after earning the EM6 or CETa certifications, go on to earn the ETA fiber optics cert.

A short-distance fiber optic link usually consists of an infrared LED, a bundle of transparent plastic fibers, and a phototransistor detector. A long-distance link uses a diode laser, a glass fiber and a PIN diode detector. Such a system can transmit data at frequencies well into the hundreds of megahertz, over many miles of cable. Copper wires can't do that.

I have more to say about fiber optics, but I want to say it in the EM6 study guide, where we'll study signal transmission in detail.

End of Chapter 10

CHAPTER 11 – EM3 REVIEW

CHAPTER		PAGE
1	Analog Signal Basics	2
2	Active Devices	4
3	Transistor Amplifiers	9
4	Opamps, DC Amps	13
5	Linear Integrated Circuits	14
6	Oscillators	17
7	Wave Shaping Circuits	19
8	Filters	20
9	Power Supplies	23
10	Optoelectronics	25
---	Charts & Graphs	26

CHAPTER 1 – ANALOG SIGNAL BASICS

SIGNAL – a wired or wireless transmission that bears information.

DIGITAL – a signal where information is represented by changes among specific voltage levels, usually two, which occur at specific time intervals. If there are small voltage or time variations (e.g. where the voltage is a little low due to loss in a long cable), the variation is not deliberate and does not represent information.

ANALOG – signal where information is represented by a continuous range of voltage and a continuous flow of time, with no steps. The important aspects of an analog signal are:

- **LEVEL**, known as amplitude. "Level" may refer to a voltage level or a power level depending on the type of signal.
- **RANGE OF LEVELS**, known as **dynamic range**.
- **SPECTRUM**, the range of frequencies present in a signal
- **IMPEDANCE**, and whether or not impedances are matched throughout the system
- **FORMAT**, including balanced or unbalanced, type of cable, type of connector etc
- **NOISE** and **DISTORTION** of the signal that can be tolerated

INTERFACE – a junction between electronic circuits where information is transferred.

DECIBEL or **dB** – a logarithmic comparison between levels. A 10:1 power ratio is 10dB; a 10:1 voltage ratio is 20dB. This is because a change in voltage also changes current, and power is voltage times current; a 10:1 voltage change results in a 100:1 power change. Decibel formulas and charts are included at the end of this review.

REFERENCE LEVELS – decibels are comparative numbers, not absolute levels. However, they can be turned into absolute levels by comparing them to specific reference levels. The most commonly used levels are:

- **0dBV** – 1 volt
- **0dBm** – 1 milliwatt
- **0dBu** – about 0.78V, which is 1mW in a 600Ω system
- **0VU** – +4dBu, 1.23V
- **0dB SPL (sound pressure level)** – an acoustic reference that corresponds to the quietest sound we can hear, measured as air pressure variations
- **0dB SIL (sound intensity level)** – an acoustic reference that corresponds to the quietest sound we can hear, measured in terms of power.

A level may be higher than its associated reference, as in +6dBV, or lower, as in -10dBV.

STANDARD LEVELS – actual levels, which differ from reference levels. Used mostly in audio.

- **Professional level**: +4dBu, 1.23V. This is the level that would read "0" on a VU meter.
- **Consumer level**: -10dBV, about 0.32V

Review page 2

SIGNAL-TO-NOISE RATIO – the level difference between desired signal and undesired noise, measured in decibels.

THERMAL NOISE – the unavoidable noise present in any system, caused by thermal agitation of atoms and molecules. Thermal noise has a hissing sound like escaping air. The amount is proportional to the square root of resistance, and also the square root of bandwidth, and to a certain extent proportional to temperature. At room temperature, a 50Ω resistance creates 10_{-9}V of noise over a bandwidth of 1Hz.

NOISE FLOOR – background noise of the system, which forms a definite lower limit to the dynamic range. This should be the thermal noise level, but there may be other kinds of noise (e.g. hum) that raise the noise floor. Usually measured in dB, as in noise = -120dBV.

CLIPPING LEVEL – the opposite extreme from the noise floor, defining the highest level the system will handle without overload. Measured in voltage, power or dB as in +20dBu.

SYSTEM DYNAMIC RANGE – the level difference between the clipping level and the noise floor, measured in dB.

SIGNAL DYNAMIC RANGE – the difference between the strongest and weakest signal. Even at the signal's weakest, the noise floor should still be well below the signal. This means a system should have a wider dynamic range than the signal itself, measured in dB.

HEADROOM – the margin between the signal at its strongest and clipping level, measured in dB.

DISTORTION – usually refers to alteration of the signal waveform, most often measured in percentage.

AUDIO – signals whose frequency content is in the range humans can hear, i.e. roughly 20Hz – 20kHz. Audio systems typically do not match impedances throughout;

RF (radio frequency) – signals whose frequency content begins above the audio band and may go as high as technology is capable of generating. Currently, the highest frequency widely used by average Joes and Jo's is around 4GHz, 4×10^9Hz.

BASEBAND VIDEO – video as it comes out of the camera. Traditional video has a frequency spectrum of 30Hz to around 4MHz. High-definition video may go up to 100MHz.

MODULATED VIDEO – video as it's transmitted over the air. In traditional analog video, still used in cable systems, each TV channel occupies a 6MHz chunk of spectrum, beginning around 50MHz for channel 2.

LINE LEVEL – a practical signal level for distribution and processing. For audio, video, and data signals, this tends to be in the neighborhood of one volt.

MICROPHONE LEVEL – the signal as it comes out of a microphone, typically 30 to 60dB lower than line level.

Review page 3

SPEAKER LEVEL – much stronger than line level. Not necessarily higher voltage, but much lower impedance, resulting in much higher current and power.

HEADPHONE LEVEL – voltage similar to line level, but lower impedance, typically 32W, thus higher power.

POTS (plain old telephone service) – the audio system used by the wired telephone system.

- **Wiring** – unshielded twisted pairs, balanced or pseudo-balanced. Grounded at the telephone company end, no connection to ground at the subscriber (user) end.
- **Talk voltage** – the DC voltage across the pair. 48V when the telephone is not in use, it drops to about 5V when the receiver is taken off the hook.
- **Audio level** – peaking at 1V with very loud speech.
- **Ring voltage** – usually 90VAC at 20Hz.
- **Rotary dialing** – done by briefly opening the circuit, equivalent to momentarily hanging up.
- **DTMF (dual tone multi-frequency)** – dialing via pairs of specific frequencies generated within the telephone.

CHAPTER 2 – ACTIVE DEVICES

ACTIVE DEVICE – a control device which performs an operation on an electrical current, boosting, modifying, processing it. Active devices are essentially valves of various kinds.

THERMIONIC – describes a device in which a hot piece of metal emits electrons into a vacuum. The control of these electrons then constitutes the active function.

SOLID STATE – describes a device in which electrons flow though a type of solid material that permits current to be controlled in a certain way, either by the construction of the material or by an external control signal.

SEMICONDUCTOR – the type of material used in solid state devices. This is a crystal, usually silicon, which is divided into several regions. Minute amounts of impurities called **dopants** are deliberately mixed into each region to provide either extra electrons or their positive equivalent, **holes**, which are places in the crystal where electrons would normally be but aren't due to the presence of a dopant atom. Holes and extra electrons can move freely through the semiconductor. A semiconductor region with free electrons is called **N-type** and a region with holes is called **P-type**. The number and arrangement of N- and P- regions determine what kind of device the semiconductor will be and what control functions it will perform.

JUNCTION – where N- and P-type regions meet. This is not a contact, in the usual sense, between two pieces of material; rather, the two regions are fused into a continuous crystal with regions that differ in their doping.

Review page 4

DIODE – an active device with two elements. In a thermionic device, the two elements are the **anode** or **plate** and the **cathode**, which may be the **filament** or a cathode indirectly heated by the filament. In a solid state device, the two elements are a P-type region which forms the anode and an N-type region which forms the cathode. A diode is a one-way valve that passes current when the anode is positive with respect to the cathode but blocks current in the other direction. This is called **rectification**. A high-power diode is usually called a **rectifier**.

FORWARD VOLTAGE – a potential across a diode where the more positive voltage is connected to the anode. When the forward voltage reaches about 0.7V, the diode conducts. Under normal conditions, this 0.7V is approximately constant at any current level within the rated limits for the diode.

REVERSE VOLTAGE – a potential across the diode where the more positive voltage is connected to the cathode, and no current flows except for a small leakage current.

REVERSE BREAKDOWN VOLTAGE – the maximum non-conducting reverse voltage for a particular diode. When the voltage exceeds this limit, the diode begins to conduct. Under normal conditions, this must be avoided because it involves a fairly high voltage and may involve high current, resulting in high power dissipation and destruction of the diode.

ZENER DIODE – a diode where the reverse breakdown characteristic has been optimized for use as a **shunt voltage regulator**, where the diode regulates by "spilling" enough current to hold the voltage down to its breakdown voltage.

SCHOTTKY DIODE – a rectifier that exhibits a lower forward voltage drop and faster turn-on and turn-off (**commutation**) times than a regular rectifier.

VARACTOR – a diode intended for variable-frequency tuning. A diode's capacitance varies with reverse voltage. A varactor is not much more than a diode where this behavior has been optimized and calibrated.

TUNNEL and **GUNN** diodes – diodes with a reversal in part of their V-I curve; in this region, increasing voltage results in decreased voltage drop. This can be used to create a very high-frequency oscillator.

TRIODE – a thermionic device with three elements. In a thermionic device, these are the anode, cathode and **control grid**. A triode is a current valve where current flow between the anode and cathode is controlled by the voltage between the grid and the cathode; thus the triode is a **voltage-controlled** valve. A triode is normally "on" – it normally conducts current – and is turned "off" by making the grid negative with respect to the cathode.

BIPOLAR JUNCTION TRANSISTOR or **BJT** – a solid state device with three elements called the **emitter**, **base** and **collector**. A BJT is a current valve where current flow between the collector and emitter is controlled by current flowing between the base and emitter; thus the BJT is a **current-controlled** valve. A BJT is normally "off" – it does not normally conduct current – and is turned "on" by making a small current flow in the base. In other words, base current is the cause and collector-emitter current is the effect.

Review page 5

BETA or β - the current gain of a BJT, which is the ratio of collector current to base current. A typical transistor might have a beta of 100, which would mean that a 1mA base current would let 100mA of current flow between collector and emitter.

NPN – a transistor where the emitter, base and collector are N-type, P-type and N-type semiconductor respectively. In an NPN transistor circuit, the emitter is at the lowest (least positive or most negative) voltage. In normal operation, the base is about 0.7V more positive than the emitter. Collector voltage is determined by the remainder of the circuit, i.e. the transistor does not control its own collector voltage, only its collector current.

PNP – the mirror image of an NPN transistor, with P-type, N-type and P-type semiconductor. In normal operation, the emitter is at the highest voltage, the base is 0.7V negative with respect to the emitter, and collector voltage is determined by the remainder of the circuit.

COMPLEMENTARY – a mirror-image relationship, as between NPN and PNP transistors. An NPN and PNP transistor designed to have as nearly identical characteristics as possible is called a **complementary pair**.

OPERATING POINT – the voltages and currents present in a circuit in the absence of signal. Where the circuit must handle a signal without altering or clipping off any part of it, it is designed so that it draws current at all times. This is called **quiescent current**. Then the current can increase and decrease to accommodate positive and negative signal alternations. Setting the operating point is also called **biasing**.

THERMAL RUNAWAY – what happens in a poorly designed circuit when a transistor's quiescent current causes heating, which causes more quiescent current, which causes more heating, leading to destruction.

SECOND BREAKDOWN – what happens when a transistor is destroyed due to concentration of collector current in a small area. Happens mostly at high voltages and high frequencies.

SAFE OPERATING AREA or **SOA –** the range of operating conditions which do not exceed a transistor's maximum voltage, current, power or SOA limits.

SATURATION – what happens when a transistor passes enough current to pull the collector voltage all the way to the emitter voltage, so there is almost no voltage across the transistor. This is desirable when a transistor is used as a high-power switch, because with little voltage drop across the transistor, little heat will be generated. On the other hand, when a transistor is saturated, i.e. fully on, it takes some time for it to turn off.

DARLINGTON TRANSISTOR – two transistors in one case whose combined beta is the product of the individual betas, typically at least 1,000.

Review page 6

FIELD EFFECT TRANSISTOR or **FET** – a solid state device with three elements called the **source, gate** and **drain**. A FET is a current valve where current flow between the drain and source is controlled by voltage between the gate and source; thus the FET is a **voltage-controlled** valve. In other words, the gate voltage is the cause and drain-source current is the effect. In normal operation, no current flows in the gate; that is, the input is an open circuit except for capacitance.

CHANNEL – the area between the source and drain which is the main current path through the device. A voltage at the gate changes the effective width of the channel, that is, the portion of the channel that will conduct current. This is how current is controlled. If the conductive part of the channel is narrowed to the point where the it disappears, current flow stops. This is called **pinch-off**.

DEPLETION MODE FET – a "normally on" FET where gate voltage must be used to turn the FET off.

ENHANCEMENT MODE FET – a "normally off" FET where gate voltage must be used to turn the FET on.

JUNCTION FET or **JFET** – a FET where the interface between the gate and the channel is a regular semiconductor junction. All junction FETs are depletion mode.

MOSFET or **metal oxide semiconductor FET** – a FET where the gate is insulated from the channel by a thin layer of metal oxide, or more recently by other materials (though the MOS designation is still used). A MOSFET can be depletion mode but the vast majority are enhancement mode.

N-CHANNEL – a FET where the source, drain and channel are N-type material and the gate is P-type material. For the input of a JFET to be an open circuit, the gate-channel junction must always be unbiased or reverse biased. Thus, an N-channel JFET must be depletion mode; it is "on" and a negative voltage is applied to the gate to turn it "off." A MOSFET, with its insulating layer between gate and channel, does not have this limitation, so a MOSFET can be enhancement or depletion mode.

P-CHANNEL – the mirror image of the N-channel FET, with source, channel and drain of P-type material and the gate N-type.

AVALANCHE DEVICE – a device that is either on or off, like a switch.

TRIGGER DIODE – a special kind of diode that is off until the voltage across it reaches a preset limit, at which point it turns on. Unlike a regular diode, which starts conducting when the reverse voltage limit is exceeded and stops conducting when the voltage falls below the limit, an "on" trigger diode is nearly a short circuit, pulling the voltage almost to zero and continuing to conduct. The only way to turn a trigger diode off is to completely remove the voltage across it.

Review page 7

SILICON CONTROLLED RECTIFIER, SCR or **THYRISTOR** – a device with an anode and cathode, which form the current path, and a trigger input to turn the SCR on. Once an SCR is on, the only way to turn it off is to remove the voltage across it. Like any rectifier, an SCR only conducts one way.

TRIAC – essentially two SCRs back to back, so it will conduct either way.

UNIJUNCTION TRANSISTOR – functionally, a trigger diode where the trigger voltage can be programmed. Obsolete, but can be imitated by a circuit with two BJTs.

THERMISTOR – a device whose resistance varies with temperature. **A positive temperature coefficient** or **PTC** means resistance rises with rising temperature. A **negative temperature coefficient** or **NTC** means resistance falls with rising temperature.

INRUSH CURRENT LIMITER – a special NTC thermistor in series with the power source. Its resistance is initially high to limit inrush current. When it heats up, its resistance falls, essentially removing it from the circuit.

SELF-RESETTING FUSE – a PTC thermistor whose resistance suddenly rises when it reaches a preset temperature.

STUD-MOUNT DEVICE – a rectifier or power transistor where a threaded machine screw end is used both to mount the device and make one of the connections. For a rectifier, the screw end is usually (though not always) the cathode. For a transistor, the screw end is the collector.

T05 – a small metal transistor housing much used in the past.

T022 – in the past, a small metal housing. Today, T022 refers to the most common plastic transistor housing, though strictly speaking many “T022” cases should be identified by different numbers. If you say “T022” everybody will know what you mean.

T03 – the most common high-power semiconductor housing in the past, still found in existing equipment.

T0220 – the most common tab-type high-power semiconductor housing today.

MICA WASHER – used to transfer heat from a semiconductor to a heat sink without making an electrical connection. Plastic and other materials are also used.

HEAT SINK GREASE – a thermally conductive grease, usually white, applied to both sides of a mica washer to encourage heat transfer.

SURFACE MOUNT or **SMT** – a component package that solders directly to the surface traces of a PC board.

JEDEC PART NUMBER – An American semiconductor numbering system beginning with the number of junctions: a diode is 1NXXX and a transistor is 2NXXX

Review page 8

ASIAN PART NUMBER – A part numbering system where:

- 2SAXXX or 2SBXXX identifies a PNP transistor
- 2SCXXX or 2SDXXX identifies an NPN transistor
- 2SJXXX identifies a P-channel FET
- 2SKXXX identifies an N-channel FET

CHAPTER 3- TRANSISTOR AMPLIFIERS

AMPLIFICATION – increasing the level (voltage, current or power) of a signal. Also called **gain**. This may be increased voltage, increased current or both. All amplifiers increase power.

ATTENUATION – decreasing the level of a signal, also called **loss**. Gain and loss may be stated as a multiple, as in 10X, or in dB, as in +20dB.

AMPLIFIER – an active circuit that makes a signal larger. Power for the increased signal comes from a DC power supply; the input signal only supplies the waveform pattern and at most a small portion of the output power.

ATTENUATOR – a circuit that makes a signal smaller. This can be a passive circuit requiring no power supply. It often takes the form of a resistive voltage divider. Also called a pad.

INVERTING – polarity of output signal is opposite from polarity of input signal.

NON-INVERTING – polarity of output signal is the same as input signal.

COMMON EMITTER or **CE** – an amplifier configuration where the emitter is part of both input and output circuits. A CE amplifier has both voltage and current gain. Inverting.

COMMON COLLECTOR or **CC** – a configuration where the collector is part of both input and output circuits. A CC amplifier has only current gain. Non-inverting.

COMMON BASE or **CB** – a configuration where the base is part of both input and output circuits. A CB amplifier has only voltage gain. Non-inverting.

QUESCENT COLLECTOR CURRENT – the preset current level that flows in the collector, when there's no AC signal present. With an NPN transistor, where the collector is more positive than the emitter, conventional current flows into the collector and out through the emitter. Actual electron flow, negative to positive, is into the emitter and out through the collector. Think of it whichever way is easier for you.

QUIESCENT BASE CURRENT – the preset current level that flows in the base, which is what causes the transistor to allow current to flow in the collector. Collector current is base current times the transistor's current gain or beta. This is done by forward-biasing the base-emitter junction. In an NPN transistor, this means the base will be about 0.7V more positive than the emitter.

Review page 9

COLLECTOR / EMITTER CURRENT RELATIONSHIP – emitter current is collector current plus base current. Since base current is typically 1% of collector current, for all practical purposes $I_E = I_C$.

DEGENERATION – a reduction in voltage gain which occurs in a CE or CB amplifier when a resistor is inserted between emitter and ground. This reduces waveform distortion and the effect of beta variations. Using ann emitter resistor, along with a voltage divider to set base voltage, is one way to obtain a stable operating point.

BYPASSING – connecting a capacitor across all or part of an emitter resistor to avoid degeneration. This is done when we want an emitter resistor for the sake of operating point stabilization, but do not want the degeneration that would otherwise come with it.

INPUT IMPEDANCE – in a CE circuit with an emitter resistor or a CC circuit, input impedance at the base is the emitter resistor times beta. Actual circuit input impedance is usually lower due to the shunting effect of bias resistors connected to the base. In a CB circuit input impedance is nonlinear and extremely low, the complex result of current and transistor current gain.

OUTPUT IMPEDANCE – in a CE or CB circuit, essentially equal to the collector resistor unless modified by negative feedback.

NEGATIVE FEEDBACK – feeding a portion of the signal from output back to input to correct waveform distortion and, in the case of DC feedback, stabilize the operating point. Negative feedback reduces gain and output impedance. Also called **degenerative feedback**.

VIRTUAL GROUND – an input whose AC impedance is near zero due to the action of negative feedback.

EMITTER FOLLOWER – another name for a common collector amplifier. Since emitter voltage is always equal to the base voltage with a relatively constant 0.7V offset, it will “follow” an AC waveform; that is, voltage gain is 1 (unity) or a little less. Input impedance is relatively high, output impedance quite low. Used when we want to drive a load that's too heavy (low Z) for the signal source. Non-inverting.

BOOTSTRAP – a method of raising input impedance in a CE or CC amplifier by coupling the AC voltage at the emitter back to the bias voltage divider.

COMMON SOURCE – a FET amplifier where the source is part of both input and output circuits, equivalent to a common emitter amplifier.

COMMON DRAIN – a FET amplifier where the drain is part of both input and output circuits.

COMMON GATE – a FET amplifier where the gate is part of both input and output circuits (little used).

Review page 10

SELF BIASING – in a depletion-mode FET amplifier, the necessary reverse voltage between the gate and source is obtained by putting a resistor between source and ground. Current flowing through the source creates a voltage across the resistor, raising the source voltage.

STAGE – when more gain is required than a one-transistor amplifier can provide, the signal can be successively boosted by one amplifier after another. Each such amplifier is called a stage.

COUPLING – connecting the output of one amplifier stage to the input of another. Since the DC levels present at input and output are often different, it may be necessary to insert a **coupling** or **blocking** capacitor. Another way of coupling is to use a transformer between stages. A third way is to design the stages so there is no DC difference between them. They can then be **direct coupled**.

DIFFERENTIAL PAIR – a 2-transistor amplifier with an inverting and a non-inverting input, also called a **long-tailed pair**.

CONSTANT-CURRENT CIRCUIT – a circuit where current flow is the same regardless of voltage across the circuit. If the circuit supplies a constant current from a positive DC source, it is known as a **current source**. If it drains a constant current to ground or a negative DC source, it is known as a **current sink**. Note that these names refer to conventional current flow.

LINEARITY – freedom from waveform distortion.

EFFICIENCY – the percentage of power drawn from the power source that can be taken as signal power at the output. An amplifier that draws 50W and puts out 25W has an efficiency of 50%. Low efficiency is undesirable in amplifiers that handle any significant amount of power, both because it wastes energy and because the wasted energy shows up as heat, which you have to get rid of if you don't want the transistors to melt.

CLASS A – an amplifier design where current flows during every part of the signal waveform. With a sinewave signal, maximum efficiency is 50% at full output. A class A amp is linear, more or less. Current draw is the same regardless of input signal.

CLASS B – an amplifier design where current flows approximately half the time. A class B amplifier can be made to handle an entire waveform by providing a complementary pair of half-amplifiers, typically an NPN and a PNP output transistor, one to handle positive alternations and one to handle negative alternations. Maximum efficiency is around 70% at full output. Current draw and power dissipation vary with signal level. A class B amp can be linear, with the exception of a waveform glitch at the **crossover** points between the positive and negative alternations. An output stage where transistors alternately share the load is called a **push-pull** output.

CLASS AB – a hybrid class where the output transistors are biased on just enough to eliminate the crossover glitch. The best class AB amps are very linear.

Review page 11

CLASS C – an amplifier design where current only flows during a small part of each cycle. Class C amps are usually RF amps where the output transistors are connected to resonant tank circuits. Efficiency in the 90% range is easily obtained. Very nonlinear; current flow comes in short, intense spikes. However, the output tank turns the spikes into a smooth sine wave.

CLASS D – an amplifier design where the output switches between positive and negative rails at a frequency much higher than the signal frequencies. The output is filtered to produce an average voltage, which is determined by the positive/negative duty cycle of the switch. If the input signal is made to change the duty cycle, the output will be a changing voltage that is a replica of the input voltage.

TURN-ON TIME – when a transistor is used as a switch, the total time elapsed from application of a turn-on signal at the base to full collector-emitter saturation.

TURN-OFF TIME – the total time elapsed from removal of a turn-on signal at the base to zero collector-emitter current.

SHOOT-THROUGH – what happens when both transistors of a complementary pair turn on; essentially, a short circuit across the power supply. In a switching circuit, this happens at the transitions because a transistor takes longer to turn off than turn on. Special timing circuitry is required to prevent this.

MILLER CAPACITANCE – an internal capacitance between collector and base, or between drain and gate, which imposes negative feedback at high frequencies. It has the effect of slowing the transistor down.

ALPHA or α – a BJT's current gain when used in a common-base amplifier. While beta may be any value, and typically falls between 50 and 400, alpha is just a tiny bit less than unity for any transistor. This is because alpha = I_C/I_E, and $I_E = I_C + I_B$. For a typical transistor, I_B is only 1% of I_C; thus I_C and I_E will be only 1% different.

f_T – the frequency at which a transistor's beta falls to unity, or, in a common base amplifier, where it falls to 0.707.

f_{MAX} – the frequency at which power gain under the best circumstances is unity; i.e. the maximum frequency of oscillation. f_{MAX} is higher than f_T.

AUDIO AMPLIFIER – a linear amplifier with a frequency range that covers all or a large part of the audio band, approximately 20Hz to 20kHz. Audio amplifiers usually operate unmatched, with a much higher input impedance than output impedance; they are not intended for a load equal to their own internal impedance.

RF AMPLIFIER – an amplifier that amplifies some portion of the radio frequency spectrum. May be narrow band or wideband, may be linear or class C. Usually operates with matched impedances of 50Ω at the input and output.

Review page 12

BASEBAND VIDEO AMPLIFIER – a linear amplifier that handles a video signal as it exists before the transmitter. Frequency range 30Hz to 4MHz for traditional video, up to 100MHz for high definition. Matched impedances, 75Ω input and output.

(MODULATED) VIDEO AMPLIFIER – an amplifier that handles analog video as it comes in from an antenna. A standard RF amplifier except that the input and output impedances are 75Ω.

LINE LEVEL AMPLIFIER – an amplifier designed to handle signals of somewhere around a volt in amplitude.

PREAMPLIFIER – an amplifier designed to boost a weak signal to line level.

POWER AMPLIFIER – an amplifier designed to boost a line level signal to the point where it will drive a speaker or other high-power load.

HEADPHONE AMPLIFIER – essentially a current amplifier, usually does not boost voltage. Designed to drive a headphone from a line level signal source. Usually designed for a 32Ω load.

BRIDGING$_1$ – attaching a load to a signal source that is so light (high impedance) that it essentially does not load the source.

BRIDGING$_2$ – attaching a load across the output of two amplifiers that are operating out of phase, so the voltage across the load is doubled compared to a single amplifier. This is how most electronically balanced outputs operate.

LIMITING AMPLIFIER – an amplifier deliberately driven into clipping to limit amplitude variations.

DC AMPLIFIER – an amplifier that is direct-coupled from input to output, so it can amplify voltage variations at low frequencies and down to DC.

CHAPTER 4 – OPERATIONAL AMPLIFIERS

OPERATIONAL AMPLIFIER, OPAMP, OP-AMP – an amplifier with a differential input and single output. Output voltage is the voltage at the non-inverting input, minus the voltage at the inverting input, times the opamp's voltage gain: $V_O = (V_N – V_I)G_V$ The native gain is extremely high, but in an actual circuit the gain is set by the feedback network.

BLACK BOX – a circuit that the user takes as a given, an electronic device that does a specific thing, but the user doesn't have to know how. An opamp is intended to be used as a black box. It is a general-purpose **gain block** with a complicated internal circuit that the user doesn't need to know anything about.

Review page 13

VOLTAGE FOLLOWER – an emitter follower transistor amplifier, or an opamp circuit with unity gain. This is obtained by connecting the inverting input directly to the output.

NON-INVERTING VOLTAGE GAIN – the ratio of the feedback resistor to the shunt resistor, plus one. A 9kΩ resistor between output and inverting input and a 1kΩ resistor between inverting input and ground would result in an voltage gain of 10X.

INVERTING VOLTAGE GAIN – the ratio of the feedback resistor to the input series resistor. When stated as a multiplier, this is a negative value due to the inversion. A 10kΩ resistor between output and inverting input and a 1kΩ resistor between inverting input and the signal source (assuming the signal source is 0Ω) would result in a voltage gain of -10X. I say "when stated as a multiplier" because, when gain is stated in dB, the gain would be positive since the output is larger than the input.

INPUT GROUND PATH – a resistive path from the + input to ground to keep the quiescent voltage from drifting.

NON-INVERTING INPUT IMPEDANCE – the opamp's input resistance in parallel with the ground path or bias resistor/resistors.

INVERTING INPUT IMPEDANCE – equal to the series resistance between the signal source and the inverting input. Due to negative feedback, the inverting input is a virtual ground.

OUTPUT IMPEDANCE – near zero in both inverting and non-inverting configurations, due to the high level of negative feedback. Where a higher output impedance is desired, a build-out resistor is connected in series with the output.

SUMMING AMPLIFIER – an amplifier whose instantaneous output voltage is the instantaneous sum of several input voltages. Virtual-ground amplifiers are popular for summing amps because, if the summing point is essentially ground, none of the signals will leak backwards into the other signals. Also, the gain for each signal is independent of the gain for other signals or how many other signals there are.

DIFFERENTIAL AMPLIFIER – with regard to opamps, a circuit that makes use of the differential nature of an opamp's inputs to create an amplifier circuit with a differential input.

OFFSET VOLTAGE – a small balancing error in the differential input circuit. In an opamp with offset, the output will be zero volts not when the inputs are both zero volts, but when there is a small difference between them. There are various ways of nulling out this offset, depending on the specific opamp used.

EM3 CHAPTER 5 – ANALOG INTEGRATED CIRCUITS

DISCRETE DEVICE – a single semiconductor device (diode, transistor, SCR) in a package.

Review page 14

INTEGRATED CIRCUIT – many semiconductor devices in a package. They may all be associated with one circuit or there may be multiple circuits, as in a package containing four opamps. Advantages of integration include lower parts cost, lower manufacturing cost, higher reliability and excellent matching between transistors.

CHIP – informal name for integrated circuit. A "chip" may actually be a small chip of silicon without leads, but more often it's a small chip of silicon connected to leads and molded into a plastic case.

PINOUT – the lead connection arrangement, i.e. which pin serves which function.

INSTRUMENTATION AMPLIFIER – a differential amplifier made with 3 opamps. A good instrumentation amp has a well balanced, high impedance input and excellent common-mode noise rejection.

COMPARATOR – a circuit used to determine which of two voltages is higher. If the non-inverting input is higher, the output will be at or near the positive supply rail. If the inverting input is higher, the output will be at or near the negative supply rail. A comparator is essentially an opamp used without negative feedback. An opamp can be used as a comparator, but a comparator cannot be used as an opamp because it is not stable with negative feedback.

POSITIVE FEEDBACK – in a comparator circuit, when the output starts to change, positive feedback encourages greater change. This has the effect of making a positive or negative transition definite, eliminating uncertainty with a noisy or slowly changing input.

HYSTERESIS – where the input voltage that causes the output to go positive is different from the input voltage that causes it to go negative. This is the effect of positive feedback.

ANALOG SWITCH – equivalent to a mechanical switch, but operated by an electronic signal. The instantaneous voltage of the signal to be switched must never stray outside the power supply voltages.

MULTIPLEXER / DEMULTIPLEXER – an analog switch that connects one lead to any selected one of several other leads. Equivalent to a mechanical rotary switch.

POTENTIOMETER – an IC consisting of a resistor tapped at many points, and circuitry to select one tap for connection to a "wiper" lead. Equivalent to a mechanical stepped potentiometer.

DRIVER IC – an IC designed to be paired with large discrete output transistors. The driver IC supplies the signal, control and monitoring functions; the discrete transistors supply the brawn.

VOLTAGE REGULATOR IC – usually takes the form of an IC that drops an unregulated DC voltage down to a lower, precisely regulated voltage.

Review page 15

MOTOR DRIVE IC – an IC that uses class D technology to develop a controlled, variable DC or AC voltage to power a motor.

TIMING CIRCUIT – an IC that performs timing functions of various types. The classic timing circuit is the 555, which every tech will run into sooner or later.

LED DRIVER IC – an IC that provides the proper drive current for a LED or a string of LEDs using class D techniques. This avoids the power waste that occurs when current is controlled by a dropping resistor.

TOUCH CONTROL and **PROXIMITY SENSOR IC'S** – use the presence of 60Hz hum or a change in capacitance to sense a touch or near-touch.

HALL EFFECT IC – another kind of proximity sensor that uses a change in a magnetic field to sense the presence or change in position of a nearby magnet.

ACCELEROMETERS and **POSITION SENSORS** – microscopic machines that use inertia or another characteristic of a mechanical sensor built right into the silicon chip to sense acceleration or position.

MEMS – short for **microelectromechanical system**. Refers to a silicon chip with a mechanical mechanism, made of silicon.

TEMPERATURE SENSOR IC – an IC where the change in a diode's 0.7V forward drop is used to detect a change in temperature.

LOGARITHMIC AMPLIFIER – an amp whose input-to-output voltage characteristic is not linear but logarithmic.

VARIABLE GAIN AMPLIFIER or **VGA** – an amplifier whose gain can be controlled by the voltage at a control pin.

AUTOMATIC GAIN CONTROL CIRCUIT – a VGA where the gain control voltage is derived from the signal itself, in such a way as to automatically adjust the gain in response to signal level changes.

COMPRESSION – feeding a signal through an AGC circuit that compresses the signal's dynamic range, reducing the degree of amplitude variation.

EXPANSION – feeding a signal through an AGC circuit that expands the signal's dynamic range, increasing the degree of amplitude variation.

COMPANSION – complementary compression and expansion, used to increase the effective dynamic range of a transmission channel.

VOLTAGE CONTROLLED OSCILLATOR or **VCO** – an oscillator whose frequency of oscillation is determined by a control voltage.

Review page 16

PHASE LOCKED LOOP or **PLL** – an IC where the phase of a VCO output is compared with the phase of an incoming signal. The result is used to generate a voltage that controlls the VCO to keep its frequency locked to the signal frequency.

SIGNAL COUPLING IC – an IC that uses light (an optocoupler) or other means to transfer a signal from one part of a circuit to another without a direct connection.

ASIC – short for **application-specific IC**, a custom chip often made for a specific manufacturer, tailored for a specific function. There are few such functions that cannot be built with standard ICs and discrete parts, but very high production volume or some other consideration may justify the development cost of a custom chip.

DIP – short for **dual inline package**, a traditional IC with a two rows of pins on opposite sides, with the pins spaced 0.1" apart.

SIP – short for **single inline package**, an IC designed to sit vertically to conserve board space, with only one row of pins.

SMT – short for **surface mount**, an IC that doesn't have pins designed to go through PC board holes, but instead is designed to sit on the PC surface, with small tabs soldered to pads on the board.

IC PIN NUMBERING – a notch or other depression indicates the "start" point. From there, looking at the top of the IC, pins are counted counter-clockwise starting from the upper left pin.

CHAPTER 6 – OSCILLATORS AND MULTIVIBRATORS

OSCILLATOR – a circuit that generates regular, repeating waves called **oscillations** at a particular frequency.

POSITIVE FEEDBACK – also called regenerative feedback, where a portion of the output of an amplifier is fed back to the noninverting input. This leads to operating point instability. A tuned circuit uses this instability to cause the voltage to go up and down at a particular frequency.

TUNED OSCILLATOR – an oscillator whose frequency is determined by a resonant circuit.

BARKHAUSEN CRITERIA – the conditions needed in a tuned oscillator for oscillation to exist :

- a positive feedback loop
- a frequency where gain around the loop is at least unity
- a zero phase shift around the loop at that frequency

HARTLEY OSCILLATOR – an oscillator using one active device, usually a transistor, and a resonant tank, with a tap in the inductor.

Review page 17

COLPITTS OSCILLATOR – a similar oscillator, but with a tapped capacitor in the tank; that is, two capacitors in series.

NON-TAPPED OSCILLATOR – where the active device is noninverting and has greater than unity gain, the tank can be a simple inductor and capacitor.

ARMSTRONG OSCILLATOR – an oscillator that uses a separate coil, called a **tickler**, to provide positive feedback via mutual inductance. Strictly speaking, in an Armstrong oscillator the tank is connected to the base or gate and the tickler coil is in the collector or drain circuit. The reverse arrangement is also called a **Meissner** oscillator.

CRYSTAL OSCILLATOR – an oscillator that uses a vibrating quartz crystal as the frequency-controlling element. Crystal oscillators can be extremely precise and stable.

RESISTANCE – CAPACITANCE or **RC OSCILLATOR** – an oscillator where the oscillation frequency is set by an RC network. Both positive and negative feedback are used. Because an RC circuit doesn't resonate, automatically producing a sine wave like an LC circuit, loop gain must be very precisely set to unity. This is achieved by an AGC function that may be as simple as an incandescent lamp.

WEIN-BRIDGE – an RC bridge network consisting of a series RC leg and a parallel RC leg, which has a broad peak in its response. The bridge forms the positive feedback part of an RC oscillator.

BRIDGED-T – an RC network that has a broad dip in its response. The network forms the negative feedback part of an RC oscillator.

PHASE SHIFT OSCILLATOR – an oscillator with an RC phase-shift feedback network between the output and inverting input. The phase shift varies with frequency; oscillation occurs at the frequency where loop phase shift = 360° (zero^{O}). Though the RC circuit is not tuned at all, the overall circuit with amplifier is sharply tuned.

BEAT FREQUENCY OSCILLATOR or **BFO** – an oscillator that uses intermodulation between two high frequencies to generate a low frequency. Its main virtue is that a small change in one of the high frequencies causes a large proportional change in the low frequency; thus it is fairly easy to design a variable frequency oscillator that covers a wide frequency range in one sweep.

MULTIVIBRATOR – an untuned oscillator where frequency is determined not by a resonant circuit but by RC time constants. An LC oscillator generates a sinewave due to a resonant tank; an RC tuned oscillator generates a sinewave via a broadly tuned RC network and a linear amplifier. A multivibrator does not generate a sine wave; the transistors saturate and its output is a square wave or a modified square wave.

ASTABLE – without stability, a circuit without a stable operating point, thus an oscillator.

Review page 18

SQUARE WAVE GENERATOR – most often, a multivibrator where the same RC components control the positive and negative alternations; thus the time constants are the same and the duty cycle is 50%.

RECTANGLE WAVE GENERATOR – a multivibrator where a diode shorts part of a resistor during half the cycle (or similar arrangement), thus generating a wave with something other than a 50% time constant.

TRIANGLE WAVE GENERATOR – a multivibrator where the output is taken from one end of the capacitor in the RC timing circuit; the result is a waveform consisting of charge-up and charge-down ramps.

SAWTOOTH WAVE GENERATOR – a triangle generator where a diode makes the ramp rise time and fall time very unequal.

TRAPEZOID WAVE GENERATOR – a triangle wave generator where the waveform peaks are clipped off, resulting in a square-like wave where rise and fall times are not instantaneous.

MULTIVIBRATOR SINE WAVE GENERATOR – a triangle wave generator where diodes or another type of nonlinear circuit squash and round off the waveform peaks. This results in a **piecewise** sine wave which is only an approximation of a true, low-distortion sine wave.

FUNCTION GENERATOR – a multivibrator-based piece of test gear that generates multiple waveforms.

MONOSTABLE MULTIVIBRATOR – also called a **one-shot**, a circuit that generates a single output pulse of specific, precise duration in response to an input pulse of any duration. In contrast to an astable circuit, a monostable circuit has one stable state, where it rests until it is "poked" by an input pulse. It then switches to its other state, but that state is unstable, so it returns to its stable state.

RANDOM NOISE GENERATOR – not an oscillator, a circuit that puts out white or pink noise. This can either originate with a noisy component, or it can be "faked" by a **pseudorandom** noise generator that's a digital circuit with an output that's a lot like noise.

CHAPTER 7 – WAVE SHAPING CIRCUITS

DIFFERENTIATOR – a circuit that extracts the change in a signal and presents it as an output; what the signal is now minus what the signal was a moment ago.

INTEGRATOR – the opposite of a differentiator, a circuit that takes a changing signal and averages it over time.

SQUARE TO TRIANGLE WAVE CONVERSION – if you feed a square wave into the input of an integrator, a triangle wave will appear at its output.

Review page 19

TRIANGLE TO SQUARE WAVE CONVERSION – if you feed a triangle wave into the input of a differentiator, a square wave will appear at its output.

CLIPPER – a circuit that limits voltage excursion in the positive or negative direction or both. This could be used to clip the peaks off a sine wave to turn it into a square wave.

PRECISION RECTIFIER – a circuit that uses an opamp to rectify an AC signal without the forward voltage drop of simple rectifiers.

CLAMP – a circuit that creates a momentary short between an AC signal line and a DC voltage reference during part of each cycle, forcing the voltage on the AC line to equal the DC reference at that moment. Often used to restore the DC component of a mixed DC+AC signal that has been lost in transmission.

RAMP GENERATOR – a circuit that creates a voltage that rises or falls at a constant rate. One type of ramp generator consists of a rectangle wave generator followed by an integrator. Another is a triangle wave generator modified to exhibit unequal rise and fall times.

CHAPTER 8 – FILTERS

FILTER – a circuit designed to boost or attenuate some frequencies as opposed to other frequencies.

FREQUENCY RESPONSE – the voltage (or power) versus frequency characteristic of a component or system. Usually presented as a graph with frequency plotted on the X-axis and signal strength in volts, watts or decibels plotted on the Y-axis. If the Y-axis is calibrated in decibels, the chart is called a **Bode** plot.

PASSBAND – the range of frequencies a particular filter passes from input to output without significant attenuation. Frequency response in this range is essentially flat.

ATTENUATION BAND – the range of frequencies a particular filter attenuates.

TRANSITION, TURNOVER, CROSSOVER, CUTOFF, or 3dB DOWN FREQUENCY – the transition frequency between passband and attenuation band. Attenuation at this frequency is 3dB compared to the passband.

ROLLOFF – the rate at which attenuation in the attenuation band increases as frequency gets farther from the turnover frequency. If the rolloff is extremely fast, so frequencies not far from the crossover frequency are essentially blocked, the attenuation band may be called a **stopband**.

ORDER – the number of reactive components in a filter. Higher-order filters with more reactive components generally have faster rolloffs.

OCTAVE – a 2:1 frequency ratio

Review page 20

6dB, 12dB, 18dB PER OCTAVE – the characteristic rolloff rates of 1st, 2nd, and 3rd order filters.

LOW PASS FILTER – a filter that passes low frequencies without significant attenuation, but increasingly attenuates frequencies above the turnover frequency.

HIGH PASS FILTER – a filter that passes high frequencies but attenuates frequencies below the turnover frequency.

BAND PASS FILTER – a filter that passes frequencies within a defined range, but attenuates frequencies above and below that range.

BAND REJECT or **BAND STOP FILTER** – a filter that attenuates frequencies within a defined range, but passes frequencies above and below that range.

ALL-PASS FILTER – a filter that passes all frequencies. Despite its flat response, this is considered a filter. An all-pass filter alters the phase relationships, shifting the phase of some frequencies relative to other frequencies.

RC FILTER – a filter made of a resistor and capacitor (or several, for higher order filters). The turnover frequency in an RC filter is the frequency where R = C.

RL FILTER – a filter made of a resistor and inductor (or several, for higher order filters). The turnover frequency in an RL filter is the frequency where R = L.

LC and **RLC FILTERS** – a filter made of an inductor and capacitor, plus some resistance (or multiples for higher order filters). An LC filter is **resonant** – it rings – and its resonant frequency is the frequency where L = C.

Q – the sharpness of tuning of an RLC filter.

D or **DAMPING** – the opposite of Q; higher Q means sharper tuning, higher D means broader tuning. D= 1/Q.

BUTTERWORTH FILTER – a 2nd order or higher filter with the flattest possible response in the passband.

CHEBYSHEV FILTER – a 2nd order or higher filter which achieves a faster initial rolloff than a Butterworth filter, at the cost of some ripple – response variation – in the passband and greater ringing on transients.

BESSEL FILTER – a 2nd order or higher filter which exhibits minimum phase delay and ringing on transients at the cost of poorly defined turnover characteristics.

ELLIPTICAL or **CAUER FILTER** – a 2nd order or higher filter that has ripple in the passband and stopband, used where an extremely fast initial rolloff is needed.

Review page 21

Butterworth, Chebyshev, Bessel and Elliptical filters can be either high- or lowpass. Passive filters of these types are RLC filters.

RLC BANDPASS FILTER BANDWIDTH – the frequency difference between the 3dB-down points on either side of the center frequency. This can be stated in Hz or in percentage of the center frequency.

SELECTIVITY – in a situation where there are signals at multiple frequencies and the purpose of the filter is to select one of them, selectivity is the degree to which a filter attenuates the other signals, stated in dB. For example, in the AM radio band, stations are spaced 10kHz apart, e.g. 990kHz, 1,000kHz, 1010kHz etc. If a filter tuned to 1,000kHz attenuates signals at 990 and 1010kHz by 20dB, then it has 20dB adjacent-channel selectivity.

CASCADING – sending a signal through several filters, one after the other.

DOUBLE-TUNED FILTER – a transformer whose mutual inductance is not great, so the primary and secondary act as independent inductors to a significant degree. Each winding has a capacitor across it, so as to act as a resonant filter. Depending on the degree of mutual inductance, a double-tuned transformer can have a flat-topped response, passing all signals within a designated passband about equally and efficiently rejecting signals outside the passband.

STAGGERED TUNING – another way of obtaining flat response over a passband. Several peaked-response filters are cascaded, but the tuning frequencies are staggered slightly.

CERAMIC and **CRYSTAL FILTERS** – filters using the piezoelectric and mechanical resonance properties of quartz crystals. These filters are precise, stable, and inexpensive.

EQUALIZATION – when the balance of frequencies in a system is upset – for example, when high frequencies are attenuated in a long cable – a filter with the opposite response may be used to restore flat overall response. This is the traditional meaning of “equalization.” Today, the term has come to mean any deliberate adjustment of a frequency response curve through filtering.

PRE-EMPHASIS – intentional boosting of high frequencies.

DE-EMPHASIS – intentional attenuation of high frequencies. If pre-emphasis is applied to a wideband signal such as audio before it is sent through a noisy transmission channel and de-emphasis is applied afterward, the result is flat response but reduced noise.

ACTIVE FILTER – a filter that includes an amplifier and feedback. An active filter can duplicate the behavior of an RLC filter without using inductors, which is often a practical benefit.

GYRATOR – an active RC+amplifier circuit that mimics an inductor.

Review page 22

SALLEN-KEY FILTER – an active filter that uses feedback to achieve a wide range of possible responses.

DIGITAL FILTER – a filter that operates on analog signals that have been converted to digital form. A digital filter can mimic an analog filter, or it may have special characteristics that cannot be achieved with an analog filter.

CHAPTER 9 – POWER SUPPLIES

POWER SUPPLY – the power source, usually DC, that supplies operating power to an electronic circuit. This can be a battery, but this chapter deals with circuits that take AC power from a power line and convert it to DC.

LINEAR POWER SUPPLY – a system made up of a transformer, rectifier, filter capacitor and, if the DC output must be smooth, a regulating component or circuit.

HALF WAVE RECTIFICATION – discarding either positive or negative AC alternations, resulting in half waves with an average DC voltage. This is done by putting a rectifier in series between the transformer output and the load.

FULL WAVE RECTIFICATION – rectification which uses both AC alternations by inverting either positive or negative alternations. This can be done with a center-tapped transformer secondary and two rectifiers, or with a 4-rectifier **bridge**.

SMOOTHING – evening out the “lumpy” DC waveform that comes from the rectifiers, also called **filtering**.

CAPACITOR FILTERING – smoothing the DC using a capacitor between the DC rail and ground. On every incoming waveform peak, the capacitor charges up; then it discharges slightly as the load draws off the charge until the next peak comes along.

INDUCTOR FILTERING – smoothing the DC using an inductor in series between the rectifiers and the load. An inductor acts like a flywheel, opposing sudden changes in current.

CHOKE – a common name for an inductor when used in a power supply.

SAG – describes what happens to the supply's DC output when an increasing load drags the voltage down.

SWINGING CHOKE – a choke whose inductance is deliberately made to vary with changes in current. This can be used to counteract voltage sag.

RIPPLE – the small residual AC component of the supply's DC output, caused by imperfect filtering by the capacitor or inductor.

Review page 23

FERRORESONANT TRANSFORMER – a transformer with a separate winding with a capacitor across it, forming a resonant tank. A resonant transformer self-regulates; its output is approximately constant despite significant input voltage or output load variance.

DECOUPLING – splitting a DC rail and inserting a filter to further reduce ripple and prevent a signal from leaking from one part of a circuit to another via the rail.

REGULATOR – a component or circuit that takes the DC output of the basic power supply, ripple and all, and converts it to a slightly lower voltage that is precisely controlled and has little or no ripple.

SHUNT REGULATOR – usually a Zener diode, which conducts whenever the voltage across it exceeds its reverse breakdown voltage.

SERIES REGULATOR – a transistor or IC which drops the poorly filtered DC down to the precise output voltage required.

LOW DROPOUT or **LDO REGULATOR** – a series regulator that will work with just a tiny difference between input and output voltages.

SWITCHING REGULATOR – the alternative to a linear regulator. Uses class D technology to produce a precisely regulated output voltage.

BUCK REGULATOR – a switching regulator whose output voltage is lower than its input voltage.

BOOST REGULATOR – a switching regulator whose output voltage is higher than its input voltage.

BOOST-BUCK REGULATOR – a switching regulator whose output is usually in the same range as the input voltage. At a given moment it may be higher or lower depending on how the input voltage varies.

SYNCHRONOUS RECTIFICATION – the use of timed switching instead of rectifiers to perform AC to DC conversion. The switches are usually MOSFETS.

INVERTER – usually, a backwards power supply that converts a comparatively low DC voltage to a higher AC voltage.

DC TO DC INVERTER – a switching regulator that converts a DC voltage to a much different DC voltage. For example, a DC inverter may be used to convert the 3V of a cellphone battery to the 24V required for the piezoelectric speaker amplifier.

UNINTERRUPTIBLE POWER SUPPLY or **UPS** – a unit that uses a battery and inverter to supply AC power when the normal AC power line fails.

Review page 24

VARIABLE FREQUENCY DRIVE or **VFD** – a unit that uses class D technology to convert 60Hz AC to AC at some other frequency. Since the speed of most AC motors is controlled by the power line frequency, a VFD is often the only practical way of obtaining a variable-speed AC motor.

BIPOLAR SUPPLY – a power supply with both positive and negative DC outputs.

POWER SEQUENCING – turning multiple supply outputs on and off in a predetermined sequence.

HOT SWAPPING – connecting or disconnecting parts of a circuit while power is on.

CHARGE PUMP – a multivibrator based circuit that uses one DC source voltage to obtain another voltage. The second voltage can be the same or opposite polarity. A diode voltage doubler is a charge pump.

CHAPTER 10 – OPTOELECTRONICS

OPTICS – the science of light.

OPTOELECTRONICS – electronic technology involving light.

PHOTORESISTOR – a light sensor whose resistance varies with light intensity.

PHOTOCONDUCTOR – a solid state device whose current conductance varies with light intensity.

PHOTOVOLTAIC DEVICE – a solid state device that generates a voltage from light.

PHOTODIODE – a light-sensitive diode that can be used either as a photoconductor or photovoltaic source.

PIN DIODE – an especially high speed photodiode.

PHOTOTRANSISTOR – a photoconductive transistor.

OPTO-FET – a photoconductive FET.

PHOTO-SCR – an SCR that is triggered by light.

SOLAR CELL – a large photovoltaic diode.

LIGHT EMITTING DIODE or **LED** – a diode that lights up when current passes through it. General-purpose LEDs can handle about 30mA of current. Some recent high-power LEDs can handle several amps.

Review page 25

WHITE LED – either a blue LED coated with an orange- white phosphor, or a combination of red, green and blue LEDs. Most LEDs have essentially unlimited life, but a phosphor-type white LED will eventually dim due to exhaustion of the phosphor.

LASER DIODE – a solid state device that emits intense coherent light.

COHERENT LIGHT – light with a spectrum so narrow that it amounts to a single frequency. Coherent light maintains in-phase wavefronts far from the light source.

LASER INTERFEROMETRY – using coherent light to make precise measurements at a distance.

OPTOCOUPLER – a device, usually in a DIP IC package, containing a light source and a light detector, usually a LED and a phototransistor. An optocoupler can be used to couple a signal from one place to another without an actual electrical connection, something like a transformer.

OPTO INTERRUPTER – an optocoupler with a space between the light source and the detector. When an object is inserted into the space, it breaks the light beam.

OPTICAL FIBER – a plastic or glass strand used to transport light from one place to another.

FIBER OPTIC LINK – a light source, optical fiber and light detector, used to optically transmit data over a considerable distance.

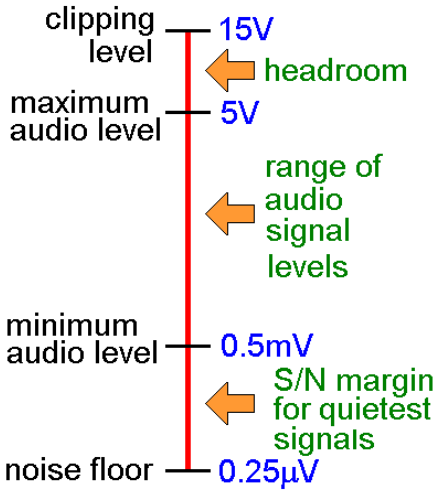
Review page 26

Decibel power formula: $\mathbf{dB = 10 \log (P_2 / P_1)}$

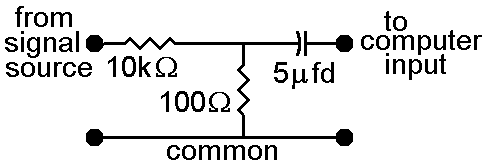
- **1dB = 26% increase or decrease in power**
- **3dB = doubling or halving of power**
- **6dB = 4x or ¼x power**
- **10dB = 10x or $^1/_{10}$x power**

Decibel voltage formula: $\mathbf{dB = 20 \log (V2 / V1)}$

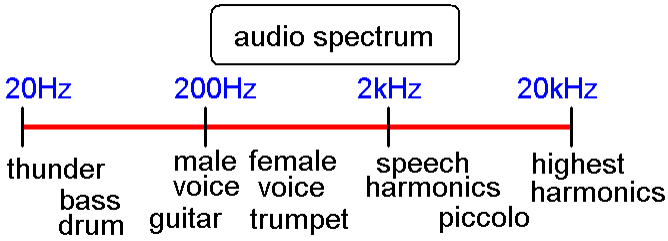
- **2dB = 26% increase or decrease in voltage**
- **3dB = 1.4 or 0.7 x voltage (square & root of 2)**
- **6dB = doubling or halving of voltage**
- **12dB = 4x or ¼x voltage**
- **20dB = 10x or $^1/_{10}$x voltage**



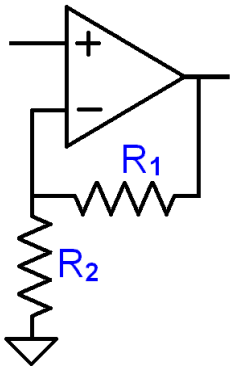
AUDIO LEVELS CHART

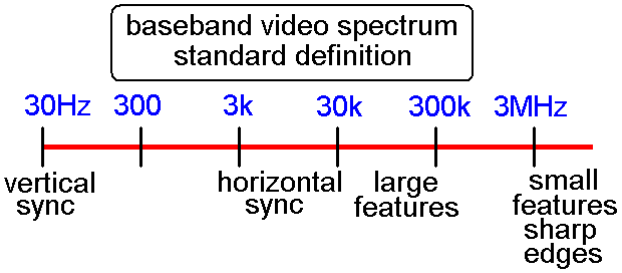


LINE-TO-MIC LEVEL ADAPTOR FOR COMPUTER INPUT

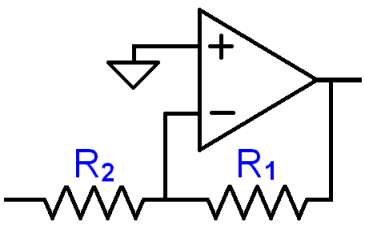


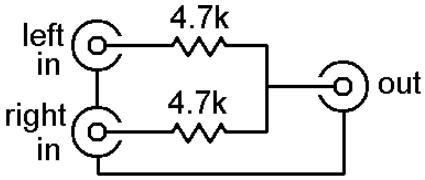
Opamp noninverting gain:
$\mathbf{G_V = 1 + R_1 / R_2}$





Opamp inverting gain:
$\mathbf{G_V = -(R1 / R2)}$

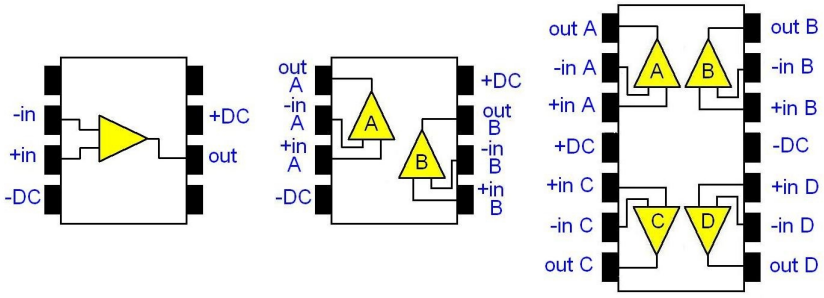




combiner cord for consumer stereo equipment

Review page 26

Most common opamp and comparator pinouts



RC filter cutoff frequency: $\mathbf{f = 10^6 / 2\pi RC}$ where C is in μf

RL filter cutoff frequency: $\mathbf{f = 10^3 R / 2\pi L}$ where L is in in mH

$\mathbf{f = 10^6 R / 2\pi L}$ where L is in μH

LC resonant frequency: $\mathbf{f = 1 / 2\pi (L^2 + C^2)^{0.5}}$

or: $\mathbf{1.59 / (LC)^{0.5}}$

Bandwidth vs. Q $\mathbf{Q = 2\ f / B}$

End of EM3 Review

EM3 PRACTICE EXAMS

Print these out and put them in a folder. Take the exams one chapter at a time, after you've read through the chapter and the review enough to be confident that you know the material. Answers for all exams are listed together, after the last exam.

Note that some chapters have a *lot* of questions. I've heard different opinions about this; some people are afraid this will be intimidating to the student. My feeling is that there is no reason to arbitrarily limit the number of questions, since this is a practice exam and you're free to not answer them all if you don't feel like it. But I'd suggest doing them all, and also going back and reviewing until your knowledge is rock solid. Besides, you've paid for this study guide, might as well get your money's worth.

Using a calculator and the charts and memory aids printed in the review is "legal."

CHAPTER 1 – ANALOG SIGNAL BASICS

1. Some fundamental aspects of a signal are:
   A. distance traveled, cost of distribution
   B. copyrighted or open source
   C. level, spectrum, impedance, format
   D. speed of transmission, whether transmitted over wires or cable

2. Signal-to-noise ratio (s/n) is:
   A. the difference between the normal signal level and the noise level in watts
   B. the difference between clipping level and the noise level in dB
   C. the difference between the normal signal level and the noise level in dB
   D. the difference between clipping level and the noise level in volts

3. Noise floor is
   A. leftover noise from the "big bang"
   B. the lowest frequency component of background noise
   C. the constant background noise level in a system
   D. the self-noise level that corresponds to a certain impedance

4. Clipping level is
   A. the level above which waveform tops and bottoms will be clipped off
   B. the level above which a standard level meter reads above zero
   C. the average signal level
   D. one volt

5. System dynamic range is:
   A. the difference between the noise level of the system and the noise inherent in the signal
   B. the difference between a "0" reading on a meter and the clipping level
   C. the difference between the noise floor and an "0" level on a meter
   D. the level difference between clipping and noise floor

6. Headroom is
   A. the difference between noise floor and clipping
   B. the difference between average level and clipping
   C. the difference between clipping level and the level that would burn out system components
   D. the difference between loud and soft moments in music

7. Source noise is
   A. noise in the preamplifier
   B. noise in the environment
   C. another expression for noise floor
   D. noise already in a signal when it comes from the source

8. If the power level of signal A is half the power level of signal B, it is _____ dB lower in level.
   A. 1dB
   B. 3dB
   C. 6dB
   D. 10dB

9. If the voltage of signal A is 10X the voltage of signal B, it is _____dB higher in level.
   A. 20
   B. 10
   C. 100
   D. 30

10. In electronics, a decade is
   A. 10Hz
   B. a 10:1 ratio
   C. a 2:1 ratio
   D. a 10:1 power gain

11. In electronics, an octave is
   A. 8Hz
   B. an 8:1 frequency ratio
   C. a 2:1 frequency ratio
   D. a 2:1 power gain

12. 0db spl is:
    A. average conversation level
    B. the quietest sound a normal ear can hear
    C. the loudest sound a human ear can tolerate
    D. one acoustic watt

13: If a sound intensity level (SIL) of 50dB is measured at 3 feet from an omnidirectional source, what will the SIL be at 12 feet?
    A. 44dB
    B. 36dB
    C. 26dB
    D. 18dB

14. 0dBm is:
    A. 1 volt
    B. 1 milliwatt
    C. 0.78 volt
    D. 1 watt

15. 0dBu is:
    A. 0dBm
    B. +4dBm
    C. about 0.78V
    D. 1 microvolt

16. 0dBV is:
    A. about 0.32V
    B. 1V
    C. the same as 0dBm
    D. 0.78V

17: 0dBW is:
    A. 1 milliwatt
    B. 1 watt
    C. +4dBm
    D. the same as 0dBV

18: -6dBW is:
    A. 0.5 watt
    B. 0.4 watt
    C. 0.1 watt
    D. 0.25 watt

19: distortion is:
    A. nonlinear change of the signal waveform
    B. a change in level
    C. a change in frequency balance
    D. a timing discrepancy

20. The “audio band” is:
    A. roughly 300Hz - 3kHz
    B. 0Hz - 20kHz
    C. roughly 20Hz - 20kHz
    D. 20kHz - 200kHz

21. Line level is:
    A. a suitable level for signal routing and transmission
    B. on the order of a volt
    C. a typical audio level in wired telephone circuits
    D. all of the above

22. Microphone level is:
    A. much lower than line level
    B. much higher impedance than line level
    C. any level in a system where Z = 150Ω
    D. equal to line voltage but lower current

23. Compared to line level, speaker level is:
    A. lower voltage and higher current
    B. higher voltage and lower current
    C. higher voltage and much higher current
    D. lower voltage and lower current

24. Headphone level is:
    A. about the same as speaker level but higher impedance
    B. about the same voltage as line level but much higher current
    C. similar to microphone level
    D. depends on the headphone

25. Combining left and right signals via a Y-cord can lead to problems because:
    A. the stereo balance is upset
    B. the two outputs fight, causing distortion
    C. it causes an impedance mismatch
    D. it causes a balanced signal to become unbalanced

26. DTMF is:
    A. a daytime transmission standard for FM stations
    B. the dual tone standard for emergency weather alert
    C. the dual tone standard for telephone dialing
    D. a method of measuring distortion using the intermodulation of two tones

27. RF is
    A. any kind of wireless transmission
    B. radio frequencies, anything above the audio band
    C. any wireless transmission of audio
    D. random frequencies, i.e. white noise

28. Baseband video is
    A. video as it comes from the camera
    B. standard definition video, 30Hz-4MHz
    C. video transmission on the lower traditional channels 2-6
    D. the output of a wireless surveillance camera

29. Modulated video is
    A. video as it comes from the camera
    B. video which has been impressed on a carrier for over-the-air or cable transmission
    C. the video output of a DVD player
    D. video with separate brightness and color signals, i.e. S-video

CHAPTER 2 – ACTIVE DEVICES

1. In an active device:
    A. an external power source is never needed
    B. a function is mechanically controlled
    C. power dissipation is always zero
    D. a function is electronically controlled

2. True/false: a discrete active device is usually an electronically controlled current valve.
    A. true
    B. false

3. Thermionic radiation:
    A. emits negatively charged particles
    B. emits positively charged particles
    C. begins at about 180° F
    D. is bidirectional; current can flow either way

4. An N-type crystal material:
    A. is slightly magnetized
    B. has a few electrons “missing” from the crystal structure
    C. has a few “extra” electrons in the crystal structure
    D. is always made of silicon

5. A diode is:
    A. a current limiting device
    B. an electrical one-way valve
    C. a measurement device
    D. a device used to increase inductance

6. A rectifier is:
    A. four diodes in one package
    B. a high-voltage diode
    C. a high-speed diode
    D. a high-current diode

7. Forward voltage drop is:
    A. the nearly constant voltage drop across a diode when it is conducting
    B. about 0.7V for silicon
    C. about 0.2V for germanium
    D. all of the above

8: Reverse breakdown voltage
    A. is the AC voltage at which a diode begins to distort the waveform
    B. is the voltage at which the diode melts
    C. is the voltage at which the diode begins to emit spurrious RF
    D. is the reverse voltage at which a diode begins to conduct

9: A zener diode:
    A. makes use of reverse breakdown
    B. is commonly used to regulate a DC voltage
    C. wastes power
    D. all of the above

10: The two most important characteristics of a schottky rectifier are: (two correct answers)
    A. small size
    B. reduced forward voltage drop
    C. increased speed
    D. low cost

11. A varactor:
    A. has much higher capacitance than a normal rectifier
    B. has conductance that varies with temperature
    C. is an electrically variable capacitor
    D. has a reverse breakdown voltage that varies with temperature

12: Varactor capacitance is highest when the DC reverse voltage across the varactor is:
    A. in the middle of its voltage range
    B. lowest
    C. as high as possible, near breakdown
    D. varactor capacitance varies with temperature, not voltage

13. The special characteristic of a tunnel or Gunn diode is:
    A. extreme high frequency operation
    B. an extremely sharp “knee” at the reverse breakdown voltage
    C. very low forward voltage drop
    D. its V/I curve has a region of negative resistance

14. The three terminals of a bipolar junction transistor (BJT) are:
    A. source, gate, drain
    B. emitter, base, collector
    C. resistor, capacitor, inductor
    D. emitter, controller, collector

15. In its natural state, a bipolar junction transistor (BJT) is:
    A. on (current flows if a voltage is present)
    B. off (no current flows)
    C. on if NPN, off if PNP
    D. low speed

16. _____ causes collector current to flow in an NPN BJT.
    A. emitter current
    B. reverse breakdown voltage
    C. base current
    D. base voltage negative with respect to emitter voltage

17. In a BJT, “beta” is:
    A. the ratio of collector current to base current
    B. the ratio of collector current to emitter current
    C. the ratio of collector voltage to base voltage
    D. the ratio of collector voltage to emitter voltage

18. In a BJT, “saturation” is:
    A. all the way off, so current = 0
    B. the state where collector voltage = base voltage
    C. maximum current the transistor can handle
    D. all the way “on,” so V_C nearly equals V_E

19. In a field-effect transistor (FET) the three terminals are:
    A. source, gate, drain
    B. emitter, base, collector
    C. source, channel, drain
    D. source, metal oxide layer, gate

20. In an N-channel MOSFET:
    A. the gate is always negative with respect to the source
    B. the gate is insulated from the channel
    C. the gate forms a PN junction with the channel
    D. gate current flows if the gate is more positive than the source

21. An insulated-gate bipolar transistor (IGBT):
    A. is another name for Darlington transistor
    B. is essentially a FET
    C. is a BJT with a FET's input characteristics
    D. is a FET with a BJT's input characteristics

22. An avalanche device:
    A. always has a separate trigger input
    B. turns fully on, all at once
    C. has built-in current limiting to prevent damage
    D. is always bidirectional (operates with AC)

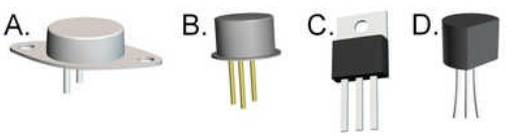
23. A trigger diode:
    A. is a 3-terminal device
    B. is “off” until a threshold is reached, then fully “on”
    C. is typically used in an RF amplifier
    D. turns off when voltage drops below the trigger threshold

24. A silicon controlled rectifier (SCR):
    A. is bidirectional (operates with AC)
    B. typically has a trigger threshold of around 30 volts
    C. turns on when it receives a current pulse at its trigger input
    D. conducts only as long as current flows in the trigger circuit

25. A triac:
    A. is essentially a bidirectional BJT
    B. is essentially a bidirectional SCR
    C. requires higher trigger current than an SCR
    D. is an SCR with an insulated (FET-like) trigger input

26. A thermistor:
    A. is a transistor whose beta changes with temperature
    B. begins conducting when a threshold temperature is reached
    C. is a heating element used to provide other components with a constant temperature environment
    D. is a resistor whose value changes with temperature

27: Which illustration below is the TO220 case?



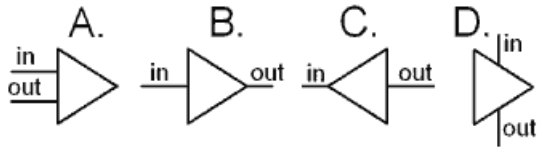
28. Which illustration above is the TO3 case?

29. The purpose of a mica washer between a transistor and heat sink is:

A. to pass heat from the transistor to the heatsink without being electrically connected
B. to decrease the Miller capacitance between collector and base
C. to decrease capacitance between the collector and heat sink
D. to accommodate physical expansion as the transistor heats up

CHAPTER 3 – TRANSISTOR AMPLIFIERS

1. The schematic symbol for an amplifier is:



2.The other common expression for "amplification" is:

A. transconductance
B. voltage transformation
C. current transformation
D. gain

3. Amplification is *not* measured as:

A. A_V, ratio of output voltage to input voltage
B. A_C, ratio of output current to input current
C. A_Z, ratio of output impedance to input impedance
D. A_P, ratio of output power to input power
A_v

4. The opposite of amplification is:

A. decline
B. attenuation or "loss"
C. diminution
D. demodulation

5. In an inverting amplifier:

A. the output waveform has the same polarity as the input
B. the output waveform has opposite polarity from the input
C. the circuit configuration is always common-base or common gate
D. the circuit configuration is always common collector or common drain

6. In a common-emitter configuration, the input voltage is applied across the _____ and the output voltage appears across the _____.

A. collector and base, collector and emitter
B. base and emitter, collector and base
C. collector and ground, collector and base
D. base and emitter, collector and base

7.In a common-collector configuration, the voltage gain is about equal to:

A. beta
B. alpha
C. 1
D. $R_C : R_E$

8. In a common-base configuration, the current gain is about equal to:

A. 1
B. beta
C. $R_C : R_E$
D. $V_C : V_B$

9. Common-emitter is the most widely used configuration because:

A. it will handle the highest frequencies
B. it provides both voltage and current gain
C. it loads the signal source least
D. its output is in phase with the input (non-inverting)

10. The most linear amplifier class is

A. class A
B. class B
C. class C
D. all amplifier classes are equally linear

11. Which configuration inverts the polarity of the signal?

A. common emitter
B. common collector
C. common base
D. all single-transistor amplifiers invert the signal

12. "Bias" is:

A. the I_C at which beta is highest
B. currents flowing in various parts of the circuit when there is no signal
C. the 0.7V (silicon) base-emitter voltage necessary to turn the transistor on
D. the ratio of collector current to emitter current

13. The reason bias is used is:

A. so current draw from the power supply will be the same under all signal conditions
B. to linearize loading of the signal source
C. so the amplifier can reproduce both positive and negative alternations
D. to improve high frequency response

14. Quiescent current is:
    A. the current level when the AC signal is at the most negative point
    B. leakage current that flows when a transistor is “off”
    C. background noise current; the current equivalent of voltage noise floor
    D. collector-emitter DC current when there is no signal

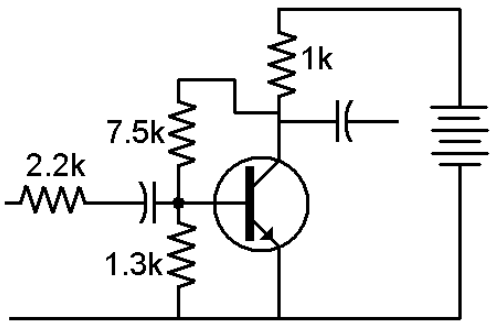
15. The “operating point” is:
    A. the normal AC signal level in a circuit
    B. quiescent voltages and currents in all parts of the circuit
    C. the temperature of a transistor when operating normally in its circuit
    D. the expected signal voltage

16. In a 1-transistor CE amplifier, it is customary to set the collector voltage to half the supply voltage (unless there is good reason to set it to some other voltage) because:
    A. it is least likely to damage the transistor
    B. it is the lowest-current operating point
    C. linearity is best
    D. it permits maximum AC output voltage

17. Coupling/blocking capacitors are used to:
    A. eliminate RF that doesn't belong on an audio signal
    B. couple AC signals into and out of an amplifier
    C. filter hum out of a power supply
    D. combine two signals

18. DC negative feedback or degeneration:
    A. stabilizes the operating point
    B. reduces waveform distortion
    C. reduces circuit gain
    D. all of the above

19. In the circuit shown below, at frequencies high enough that the capacitor does not affect operation, input impedance is a little over: (hint – should require no calculation)



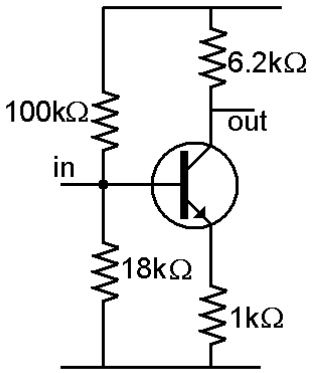
    A. 3.5kΩ
    B. 2.2kΩ
    C. 9.7kΩ
    D. 500Ω

20. A bootstrap circuit:
    A. improves high frequency response
    B. reduces loading on the signal source
    C. lowers waveform distortion
    D. lowers noise

21. In the absence of negative feedback, the output impedance of a CE or CB amplifier is approximately equal to:
    A. the collector resistor
    B. supply rail voltage divided by collector current
    C. zero
    D. the input iimpedance

22. Using a capacitor to bypass the emitter resistor in a CE circuit:
    A. improves low frequency response
    B. reduces waveform distortion
    C. reduces or eliminates AC degeneration
    D. reduces quiescent current

23. In the circuit at right, input impedance is approximately _____and gain is approximately _____. (Assume beta = 100)



    A. 18kΩ, 7.2x
    B. 100kΩ, 6x
    C. 13kΩ, 100x
    D. 13kΩ, 6x

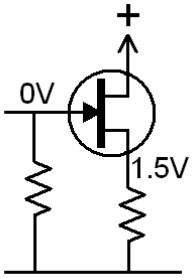
24. A common-collector amplifier is called a “voltage follower” because:
    A. the AC output voltage follows the input voltage
    B. the output waveform is slightly delayed
    C. output current does not follow input current
    D. the DC emitter voltage is the same as the DC base voltage

25. In an N-channel JFET amplifier, the gate is always _____ than the source in normal use.
    A. more positive than
    B. more negative than
    C. lower impedance than
    D. forward biased relative to

26. In an N-channel enhancement-mode MOSFET amplifier, the gate is normally _____ the source.
    A. more positive than
    B. more negative than
    C. equal in voltage to
    D. may be either A or B

27. The DC input resistance to a FET gate is:
    A. essentially zero (virtual ground)
    B. source resistance times current gain
    C. essentially infinite
    D. equal to the drain resistor

28. In the circuit shown, the bias voltage at the source is developed by the quiescent drain-source current. This is called _____ bias.
    A. degeneration
    B. self
    C. fixed
    D. passive



29. When we send a signal through one amplifier after another as in the illustration, we call each amplifier a _____.
    A. cascade
    B. booster
    C. benchmark
    D. stage

30. Three types of interstage coupling are:
    A. capacitor, transformer, direct
    B. capacitive, conductive, impedance
    C. passive, active, transconductive
    D. series, parallel, series-parallel

31. The AC output of a differential amplifier is equal to:
    A. the supply rail voltage minus collector/drain voltage
    B. the voltage difference across the inputs times the collector resistor
    C. the voltage difference across the inputs times A_V
    D. the voltage difference across the inputs times supply rail voltage

32. In the positive current convention, a current source:
    A. passes a predetermined amount of current from the positive supply, regardless of voltage
    B. has a predetermined output voltage, regardless of current
    C. passes a predetermined amount of current to ground, regardless of voltage
    D. is another name for a power supply

33. In the positive current convention, a current sink:
    A. passes a predetermined amount of current from the positive supply, regardless of voltage
    B. has a predetermined output voltage, regardless of current
    C. passes a predetermined amount of current to ground, regardless of voltage
    D. is another name for a heat sink

34. In a class A amplifier, current flows:
    A. 100% of the time
    B. slightly more than 50% of the time
    C. slightly less than 50% of the time
    D. depends on the application

35. In a class B push-pull amplifier, current flows:
    A. 100% of the time
    B. 70% of the time
    C. about 50% of the time in each half
    D. less than 20% of the time

36. In a class AB push-pull amplifier:
    A. both halves are biased off
    B. at least a small amount of current always flows
    C. current flows less than 20% of the time
    D. the output is balanced

37. In a class C amplifier, current flows:
    A. significantly less than 50% of the time
    B. more than 80% of the time
    C. exactly half the time
    D. 100% of the time

38. The least efficient class is:
    A. class B
    B. class AB
    C. class A
    D. class C

39. In a push pull amplifier:
    A. positive and negative alternations are handled by different parts of the amplifier
    B. the input and output are reversed half the time
    C. there is never any distortion at the waveform zero-crossing
    D. larger output transistors are required than in a class A amplifier of the same power

40. Crossover distortion is:
    A. a waveform distortion at peaks
    B. a waveform distortion that momentarily occurs when the signal level changes
    C. a waveform distortion at the zero-crossing
    D. a program distortion that occurs when a country music station plays a rock tune

41. In a class D push-pull amplifier:
    A. both halves conduct at all times
    B. the waveform as seen on an oscilloscope connected to the output transistors is exactly the same as the signal waveform
    C. the switching frequency must be lower than any frequency component of the signal
    D. the output switches rapidly between rails

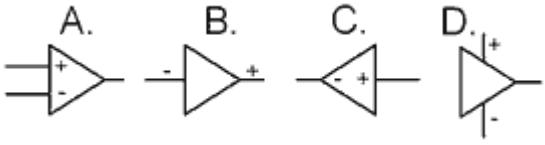
42. In audio systems, impedances are:
    A. usually not matched, except in telephone systems
    B. usually matched, except in telephone systems
    C. always 600 ohms
    D. irrelevant

43. In RF, video and data systems, impedances are:
    A. seldom matched
    B. irrelevant
    C. almost always 75 ohms
    D. almost always matched

CHAPTER 4 – OPERATIONAL AMPLIFIERS

1. An operational amplifier or op-amp:
    A. is an amplifier whose behavior is controlled by other components
    B. is an amplifier with a differential input and single output
    C. may become unstable if connected to a capacitive load
    D. all of the above

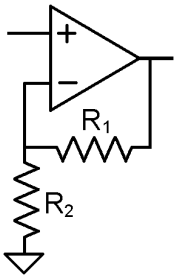
2. The schematic for an op-amp is:



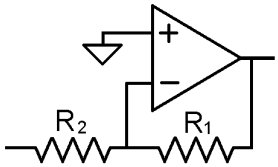
3. The formula for op-amp gain is:
    A. $1\text{-}A_V(V_{+INPUT} - V\text{-}_{INPUT})$
    B. $A_V(V_{+\,INPUT} - V_{-INPUT})$
    C. $(V_{+INPUT} - V_{-INPUT}) / A_V$
    D. $A_V(V_{+INPUT} / V_{-INPUT})$

4. The reason an op-amp's native gain is very high is:
    A. so we can build a very high gain amp if we want to
    B. so we'll have some gain left at high frequencies, where the response curve is falling
    C. high gain amps have low noise
    D. to insure that circuit behavior is controlled by other components

5. Non-inverting op-amp circuit gain is given by the equation:
    A. $A_V = 1 + (R_1 / R_2)$
    B. $A_V = (R_1 / R_2)$
    C. $A_V = - (R_1 / R_2)$
    D. $A_V = 1 + (R_1 - R_2)$

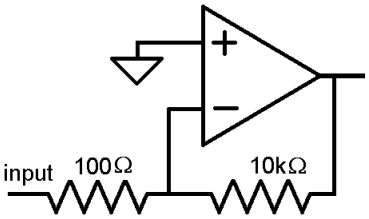


6. Inverting op-amp circuit gain is given by the equation:
    A. $- R_2 / R_1$
    B. $(R_2 - R_1)^2$
    C. $- R_1 / R_2$
    D. $R_1 - R_2$



7. If a FET-input opamp is used in noninverting configuration, input impedance is:
    A. essentially zero
    B. determined by other components
    C. determined by circuit gain
    D. determined by opamp gain

8. In the inverting circuit shown below, input impedance is about:



    A. 100Ω
    B. 10kΩ
    C. 10,100Ω
    D. 99Ω

9. When used in an actual circuit, opamp output impedance is:
    A. the feedback resistor divided by circuit gain
    B. the feedback resistor divided by opamp gain
    C. close to zero
    D. close to infinity

10. "Offset" is:
    A. a constant DC voltage error
    B. gain error
    C. time delay between input and output
    D. the difference between circuit gain when working into an open circuit and gain when loaded

11. The power supply rails should be bypassed to ground near the IC to prevent:
    A. instability
    B. oscillation
    C. waveform distortion
    D. all of the above

CHAPTER 5 – ANALOG INTEGRATED CIRCUITS

1. An advantage of integrated circuits is:
    A. lower final cost
    B. improved transistor matching
    C. high performance
    D. all of the above

2. The connection pattern of an IC is called the:
    A. pinout
    B. interface
    C. graphic design
    D. impedance match

3. An instrumentation amplifier:
    A. has unbalanced inputs and outputs
    B. uses one high precision opamp
    C. uses three opamps
    D. has a gain of slightly less than unity

4. A comparator:
    A. shows which of two frequencies is higher
    B. shows which of two voltages is higher
    C. shows which of two DC power rails has more noise on it
    D. shows which of two resistances is higher

5. A comparator with positive feedback has:
    A. hysteresis
    B. oscillation
    C. high stability
    D. slow speed

6. The process of quickly switching among several signals so as to combine them on one wire is called:
    A. selectivity
    B. mixing
    C. compression
    D. multiplexing

7. The process of quickly switching the multiplexed output so as to redistribute the individual signals is called:
    A. spreading
    B. fanout
    C. distribution
    D. demultiplexing

8. A potentiometer approximated by many fixed resistors, with a means of selecting among voltage divider ratios, is a:
    A. comparator
    B. stepped potentiometer
    C. selector switch
    D. rheostat

9. The primary requirement for a baseband video amplifier is:
    A. extremely wide, flat frequency response
    B. 50-ohm input and output impedances
    C. exceptionally low distortion
    D. gain stability within 0.3%

10. A linear voltage regulator IC is essentially:
    A. a high power shunt, like a big zener
    B. a high-power opamp in a feedback loop
    C. a solid state imitation of an inductor
    D. a transistor

11. An H-bridge is commonly used:
    A. to precisely measure resistances
    B. to increase amplifier gain
    C. to drive DC motors
    D. to compare voltages

12. A 555 IC is used to:
    A. provide a square wave signal
    B. provide a single pulse of timed duration
    C. provide linear amplification
    D. A and B above

13. The advantage of a switching LED driver IC over a dropping resistor is:
    A. lower cost
    B. less power wasted
    C. longer life
    D. less RF interference radiated

14. A Hall effect IC:
    A. is controlled by light
    B. is controlled by an external magnetic field
    C. is a temperature sensor
    D. is a vibration sensor

15. “MEMS” stands for:
    A. micro-electromechanical system
    B. micro-electromotive system
    C. magnetic / electric mechanical system
    D. micro - electric motor servo

16. A logarithmic amplifier:
    A. exhibits gain that varies with logarithmic frequency
    B. has a logarithmic (i.e. nonlinear) input impedance
    C. converts a linearly varying voltage to a logarithmically varying voltage
    D. has low harmonic distortion

17. “VGA” means:
    A. voltage gated amplifier
    B. variable gain amplifier
    C. variable gain attenuator
    D. variable gain actuator

18. “AGC” means”
    A. automotive gain control
    B. analog gain control
    C. automatic gating circuit
    D. automatic gain control

19. Analog signal “compression” means:
    A. varying gain to reduce level variations
    B. removing pauses to shorten a program
    C. reducing the bandwidth of a signal so it can be transmitted over a telephone circuit
    D. clipping off excessive signal peaks

20. Analog signal “expansion” means:
    A. the same as “amplification”
    B. varying gain to increase level variations
    C. distribution to a wide audience via the web
    D. improving the low and high frequency response of a system

21. “Compansion” means:
    A. a control signal that travels along with a program signal
    B. complementary compression and expansion
    C. compression and equalization
    D. decompression and equalization

22. “PLL” means:
    A. power locked loop
    B. phase locked loop
    C. push-pull
    D. power linked loop

23. The PLL operates by:
    A. clipping the incoming signal
    B. phase-shifting the incoming signal
    C. synchronizing an oscillator with the signal
    D. changing the frequency of an incoming signal

24. An opto-isolator or opto-coupler is:
    A. a pair of modulated LEDs, one to send each way (duplex)
    B. a pairing of a light source and light detector in one package
    C. a light source and light detector coupled via an optical fiber
    D. a device that isolates one wavelength from all others

25. “ASIC” stands for:
    A. American standard integrated circuit
    B. analog system integrated circuit
    C. application-specific integrated circuit
    D. American Standard for Internet Interface

26. A DIP package consists of:
    A. a square with pins on all four sides
    B. two parallel rows of through-hole pins on 0.1” centers
    C. two rows of surface-mount pins on 0.1” centers
    D. three pins and a thermal tab, i.e. TO220

27. IC pins are numbered (looking at the IC from the top):
    A. varies by manufacturer
    B. top to bottom left, then top to bottom right
    C. clockwise from upper right
    D. counterclockwise from upper left

28. The starting position for pin numbering is identified by:
    A. a “1” printed on the case
    B. a notch, depression or 45° angle
    C. a larger pin
D. a longer pin

CHAPTER 6 - OSCILLATORS

1. An oscillator:
    A. creates white noise
    B. creates a continuous signal at a particular frequency
    C. creates a damped oscillation like a bell
    D. selects one signal frequency from among many

2. The purpose of positive feedback in an oscillator is to:

A. achieve high voltage stability
B. assure a precise oscillation frequency
C. maintain oscillation
D. reduce waveform distortion

3. Another common expression for positive feedback is:

A. regenerative feedback
B. recirculating feedback
C. supportive feedback
D. stabilizing feedback

4. The necessary conditions for self-sustaining oscillation are called the:

A. Bode criteria
B. Hewlett-Packard criteria
C. Barkhausen criteria
D. deForest criteria

5. Continuous waveform generators are usually divided into _____ and _____.

A. AF and RF
B. tuned oscillators and multivibrators
C. line operated and battery operated
D. positive and negative feedback

6. Another name for a parallel LC tuned circuit is a:

A. selector
B. gyrator
C. beat frequency oscillator
D. tank

7. In a Hartley LC oscillator, the _____ is tapped.

A. tank's inductor
B. tank's capacitor
C. emitter resistor
D. power supply

8. In a Colpitts LC oscillator, the _____ is tapped.

A. load resistor
B. tank's inductor
C. tank's capacitor
D. power supply

9. In an LC oscillator, halving L or C changes the frequency by a factor of:

A. about 2:1
B. about 1.4:1
C. about 0.7:1
D. about 0.5:1

10. The Armstrong oscillator gets positive feedback via a _____ coil.

A. feeder
B. coupler
C. tickler
D. contributor

11. The principal advantage of a crystal oscillator is:

A. small size
B. low price
C. low noise
D. frequency stability

12. At resonance, a crystal is _____.

A. capacitive
B. inductive
C. resistive
D. nonlinear

13. In an RC oscillator, halving R or C changes the frequency by:

A. 1.4X
B. 2X
C. 3X
D. 4X

14. In a Wein-bridge oscillator, the frequency-selective circuit is in the _____ loop.

A. positive
B. negative
C. ground
D. input

15. In a bridged-T oscillator, the frequency-selective circuit is in the _____ loop.

A. ground
B. positive
C. negative
D. output

16. True or false: in a phase-shift oscillator, there is no tuned circuit and frequency is controlled by a phase shift network.

A. true
B. false

17. If 1,000kHz and 1,010kHz sinewaves are mixed (intermodulated), the combined output will contain the following frequencies:

A. 10, 1,010, 2,010 and 2,020 kHz
B. 10, 1,000, 1,010 and 2,010 kHz
C. 10, 500, 990, 1,010kHz
D. cannot be determined from given information

18. An oscillator that works by combining two frequencies is called a/an:
    A. RLC oscillator
    B. beat frequency oscillator
    C. frequency multiplier
    D. extraction oscillator

19. A multivibrator's oscillation frequency is determined by:
    A. an RC time constant
    B. a tuned circuit
    C. a phase shift circuit
    D. a crystal

20. A multivibrator can be designed to put out a variety of waveforms, but generally can *not* put out:
    A. a triangle wave
    B. a 90% duty cycle rectangle wave
    C. a clean sine wave
    D. a sawtooth wave

21. True or false: it is easy to design a multivibrator with very high frequency stability.
    A. true
    B. false

22. The difference between a square wave and a rectangle wave or pulse train is:
    A. peak voltage
    B. duty cycle
    C. frequency
    D. current

23. The difference between a triangle wave and a sawtooth wave is:
    A. peak voltage
    B. average voltage
    C. duty cycle
    D. DC offset voltage

24. “Piecewise” approximation of a sine wave is typically achieved by:
    A. squashing the peaks of a triangle wave
    B. assembling parts of several sine waves
    C. highpass filtering a triangle wave
    D. DC level-shifting a triangle wave

25. A function generator:
    A. can create a variety of waveforms
    B. provides the purest sinewave
    C. always requires 117VAC power
    D. always has a balanced output

26. A trapezoid wave is:
    A. a cross between square and sawtooth waves
    B. a square wave with increased rise and fall times
    C. a sine wave shifted 90°
    D. an inverted square wave

27. A monostable multivibrator:
    A. puts out a continuous string of very short pulses
    B. puts out a continuous sinewave
    C. is stable in either condition (high/low)
    D. puts out a single pulse of specific length

28. Another name for monostable multivibrator is:
    A. latch
    B. one-shot
    C. hysteresis
    D. regenerator

29. The output of a white-noise generator:
    A. is random
    B. contains energy across a wide frequency spectrum
    C. can be obtained from a reverse-biased transistor junction
    D. all of the above

CHAPTER 7 – WAVE SHAPING CIRCUITS

1. Differentiation is:
    A. finding the difference between two DC voltages
    B. the process of deriving change vs. time information about a voltage
    C. finding the difference between two frequencies
    D. finding the difference between two voltages

2. If a square wave is fed into a differentiator, the output will be:
    A. a triangle wave
    B. alternating positive and negative pulses
    C. a square wave with a DC component
    D. a sawtooth wave

3. If a sine wave is fed into a differentiator, the output will be:
    A. a sine wave with a DC component
    B. a sine wave with shifted phase
    C. a triangle wave
    D. depends on frequency

4. The sinewave frequency response of a differentiator is _____ frequency
   A. inversely proportional to
   B. proportional to
   C. unrelated to
   D. proportional to the square of

5. The formula for the output of a differentiator with a basic voltage gain of 1 is:
   A. $V_{OUT} = \Delta V_{IN} / \Delta_{TIME}$
   B. $V_{OUT} = \Delta V_{IN} \times \Delta_{TIME}$
   C. $V_{OUT} = \Delta_{TIME} / \Delta_{VIN}$
   D. $V_{OUT} = \Delta_{TIME} \times \Delta V_{IN}$

6. The output of a differentiator is called the _____of the input.
   A. distillation
   B. derivative
   C. deduction
   D. diminutive

7. A “step” voltage change is:
   A. a single, short pulse
   B. an integer change in DC voltage, e.g. from 6V to 7V (but not 6.2V to 6.6V)
   C. a slow DC voltage change
   D. a single, sudden change of DC voltage

8. Integration is:
   A. combining two or more signals
   B. a combination of DC and AC
   C. the average of voltage over time
   D. cancellation of a DC offset

9. If a square wave is fed into an integrator, the output will be:
   A. a series of pulses
   B. a triangle wave
   C. a sine wave
   D. a distorted sine wave

10. If a sine wave is fed into an integrator, the output will be:
   A. a sinewave with shifted phase
   B. a triangle wave
   C. a DC voltage with no AC component
   D. a sine wave at a lower frequency

11. The sinewave response of an integrator is _____ to frequency.
   A. logarithmic
   B. proportional
   C. inversely proportional
   D. unrelated

12. When the input to an integrator is a constant DC voltage, the formula for the output is:
   A. $V_{OUT} = V_{IN} / \Delta_{TIME}$
   B. $V_{OUT} = \Delta_{TIME} / V_{IN}$
   C. $V_{OUT} = V_{IN} \times \Delta_{TIME}$
   D. $V_{OUT} = V_{IN} + \Delta_{TIME}$

13. The output of an integrator is called the _____ of the input.
   A. interval
   B. integration
   C. integral
   D. intermix

14. A clipping circuit limits the _____ of an input waveform.
   A. positive and/or negative voltage excursion
   B. highest frequency component
   C. distortion
   D. duration

15. If a large sine or triangle wave is fed into a clipper, the output will be:
   A. a triangle wave
   B. approximately a square wave
   C. a sawtooth wave
   D. DC

16. A clamping circuit sets the _____ of an input waveform.
   A. duration
   B. frequency spectrum
   C. most positive or negative voltage peak
   D. noise floor

17. One use for a clamping circuit is:
   A. to remove distortion from a sine wave
   B. to set the DC level of a baseband video signal
   C. to change the frequency of a signal
   D. to change the impedance of a signal

18. A ramp generator creates:
   A. a sine wave with smoothly rising frequency
   B. a smoothly rising or falling voltage
   C. smoothly rising gain in an amplifier
   D. smoothly rising impedance

CHAPTER 8 - FILTERS

1. Most broadly defined, a filter:
   A. removes 60Hz hum from a signal
   B. removes noise from a signal
   C. treats some frequencies differently than others
   D. changes the high / low frequency balance in an audio system

2. A graph of a filter's frequency response usually shows:
   A. its voltage response vs. frequency in dB
   B. its impedance vs. frequency in Ohms
   C. its voltage response vs. frequency in volts
   D. its phase shift vs. frequency in degrees

3. A Bode plot is:
   A. a graph whose x-axis is logarithmic frequency and y-axis is voltage in dB
   B. a circular graph showing amplitude and phase
   C. a graph of impedance vs. frequency
   D. a graph of distortion vs. frequency

4. True or false: the impedance of the source and load circuitry on either side of a filter has a great effect on the filter's behavior.
   A. true
   B. false

5. A filter's “passband” is the range of frequencies:
   A. where response is greatest
   B. where impedance is purely resistive
   C. where response is flat
   D. near the 3dB down point

6. A filter's “attenuation band” is the range of frequencies:
   A. where impedance is mostly reactive
   B. where response is attenuated relative to the passband
   C. where response is zero
   D. near the 3dB down point

7. “Stop band” means:
   A. another name for the attenuation band in filters with a steep rolloff
   B. the band starting at half or twice the turnover frequency for high and low pass, respectively
   C. frequencies where filter attenuation is infinite
   D. frequencies where input impedance is infinite (i.e. open circuit)

8. “Turnover frequency,” “transition frequency,” “cutoff frequency” refer to:
   A. the demarcations between bass, midrange and treble in an audio system
   B. the transition frequency between passband and attenuation band
   C. the demarcations between adjacent RF bands, e.g. VHF and UHF
   D. the frequency where one filter stage hands off to another

9. “Rolloff” means:
   A. the rate of attenuation increase vs. frequency outside the passband
   B. the slight attenuation in the passband near the transition frequency
   C. the sharpness of the curve at the transition frequency (the “knee”)
   D. impedance change in the attenuation band

10. The turnover frequency is generally considered to be the frequency where response is ____dB down compared to the passband.
   A. 1
   B. 3
   C. 6
   D. 10

11. An octave is:
   A. a 1.4:1 frequency ratio
   B. a 10:1 frequency ratio
   C. a 2:1 frequency ratio
   D. a 2:1 impedance ratio

12. The rolloff of a filter containing a single resistor and a single reactive element exhibits a slope of:
   A. 6dB per octave
   B. 3dB per octave
   C. 12dB per octave
   D. need more information to answer

13. The filter above would be called a _____ filter.
   A. first order
   B. first class
   C. second order
   D. single ended

14. The turnover frequency is the frequency where the resistance and reactance:
   A. cancel each other
   B. are 180° out of phase
   C. sum to equal the source and load impedance
   D. are equal to each other

15. The formula for finding the turnover frequency of an RC circuit is (capacitance in microfarads):
    A. $f = 10^6 \times 2\pi RC$
    B. $f = 10^6 / 2\pi RC$
    C. $f = 2\pi / 10^6 RC$
    D. $f = 10^6 \times RC / 2\pi$

16. A lowpass filter:
    A. is an open circuit at high frequencies
    B. attenuates signals above the turnover frequency
    C. attenuates signals below the turnover frequency
    D. advances the phase of the signal

17. A highpass filter:
    A. is a short circuit at low frequencies
    B. attenuates signals above the turnover frequency
    C. attenuates signals below the turnover frequency
    D. reverses the phase of the signal

18. A bandpass filter:
    A. always has a Q greater than 1
    B. attenuates frequencies above and below a defined passband
    C. is always a single RLC tank
    D. has nearly infinite impedance outside the passband

19. A bandstop or notch filter:
    A. always has a Q less than 1
    B. has near-zero impedance at the notch center frequency
    C. passes all frequencies except those within a defined attenuation band
    D. is a parallel connected LC tank between the signal wire and ground

20. An RLC filter combines:
    A. magnetic reluctance, inductance, and capacitance
    B. inductance, resistance and remanence
    C. inductance, capacitance and resistance
    D. resistance, capacitance and rectification

21. An RLC filter is a _____ order filter.
    A. 1st
    B. 2nd or higher
    C. 3rd or higher
    D. 4th or higher

22. The slope of a 2nd order high- or low-pass filter is:
    A. 6dB per octave
    B. 9dB per octave
    C. 12dB per octave
    D. 18dB per octave

23. The turnover or resonant frequency of an RLC filter is found using the formula:
    A. $f = 1.59LC$
    B. $f = 1.59\ (LC)^{0.5}$
    C. $f = LC / 1.59$
    D. $f = 1.59 / (LC)^2$

24. Damping:
    A. reduces the Q of a filter
    B. reduces the response of an RLC at and near resonance
    C. is inversely proportional to Q
    D. all of the above

25. True or false: at the turnover frequency, the output voltage an RLC low pass filter with a damping factor of 0.3 will be greater than its passband voltage.
    A. true
    B. false

26. “Critically damped” means:
    A. having a Q of 0.5 or lower
    B. exhibiting as flat a passband as possible consistent with good transient response
    C. having a damping factor of 0.5 or lower
    D. the 3dB down point is at the resonant frequency

27. “Ringing” means:
    A. generation of spurious harmonics
    B. generation of damped oscillations in response to a transient
    C. an audio filter with a metallic sound
    D. a bandpass filter that responds to harmonics and subharmonics

28. A Butterworth filter has:
    A. flattest possible response in passband
    B. no ringing on transients
    C. constant impedance within passband
    D. constant phase within passband

29. Compared to a Butterworth filter, a Chebyshev filter provides:
    A. wider range of acceptable source and load impedances
    B. smaller size for a given performance
    C. less phase shift
    D. faster initial rolloff outside the passband

30. The drawback of a Chebyshev filter is: (two correct answers)
    A. much higher cost
    B. ripple in the passband
    C. can pick up 60Hz hum
    D. greater ringing than Butterworth filter

31. Compared to a Butterworth or Chebyshev filter, a Bessel or Thompson filter offers: (two correct answers)
    A. no ringing
    B. immune to 60Hz hum
    C. smaller inductors
    D. least phase shift in passband

32. An elliptical or Cauer filter achieves:
    A. conversion between balanced and unbalanced circuitry
    B. zero phase shift
    C. very fast rolloff and a dependable minimum attenuation across the stopband
    D. zero latency in a cellphone system

33. "Bandwidth" of an RLC bandpass filter is:
    A. one octave
    B. the frequency range between -3dB points
    C. the same as Q
    D. the frequency range between -20dB points

34. Bandwidth is related to Q by the following formula:
    A. $Q = 2F_R / B$
    B. $B = 2QF_R$
    C. $Q = 2BF_R$
    D. $B = 2Q / F_R$

35. Selectivity is:
    A. the amplification of signals within the passband
    B. the degree of rejection of signals outside the passband
    C. the bandwidth at the -6dB points
    D. the range over which a variable-frequency tank can be tuned

36. A flat-topped bandpass filter characteristic, with essentially equal response over a narrow band and high rejection of frequencies outside the band, can be obtained by:
    A. stagger-tuning two tanks
    B. partial inductive coupling between two tanks
    C. tuning one tank to 2X the frequency of another
    D. A and B above

37. A ceramic or crystal filter:
    A. has a very wide bandwidth
    B. uses mechanical resonance
    C. is a short circuit at DC
    D. can be electrically tuned over a wide range

38. Broadly defined, equalization is:
    A. another term for impedance matching
    B. deliberate alteration of frequency response
    C. correction of a frequency imbalance
    D. adjusting two signal levels so they match

39. Pre-emphasis is:
    A. high-frequency boost before transmission
    B. amplification before transmission
    C. selective amplification of the speech range, 300Hz to 3kHz
    D. automatically boosting the lower-level parts of a signal

40. De-emphasis is:
    A. attenuation after reception
    B. high-frequency attenuation after reception
    C. automatically attenuating the lower-level parts of a signal
    D. filtering out low frequencies

41. An active filter is:
    A. high Q
    B. expensive
    C. a filter whose operation is based on an active component
    D. a filter whose response varies in real time

42. "Gyrator" is generally taken to mean:
    A. a phase shifter
    B. a way to imitate an inductor
    C. a filter whose phase changes radically with frequency, e.g. Chebyshev
    D. another name for oscillator

CHAPTER 9 – POWER SUPPLIES

1. "Power supply" generally means:
    A. generating plant
    B. a unit that converts AC mains power to the DC necessary to run electronic circuitry
    C. the energy stored in a capacitor
    D. a battery

2. The four parts of a typical DC power supply are:
    A. case, power cord, on/off switch, fuse
    B. transformer, rectifier, filter, regulator
    C. transformer, rectifier, trigger diode, inductor
    D. rectifier, capacitor, regulator, pilot lamp

3. Half wave rectification:
   A. results in 60Hz ripple on the DC output
   B. results in 120Hz ripple on the DC output
   C. recharges the filter capacitor twice each cycle
   D. produces least ripple

4. Full-wave rectification using a center-tapped transformer secondary requires __ rectifier/s.
   A. 1
   B. 2
   C. 4
   D. 6

5. Full-wave rectification using a non-tapped transformer secondary requires _____ rectifiers.
   A. 2
   B. 3
   C. 4
   D. 6

6. The rectifier circuit referred to in question 5 is called a _____ rectifier.
   A. square
   B. bridge
   C. cross-connected
   D. filtered

7. The voltage drop across a silicon bridge rectifier is approximately:
   A. 1.4V
   B. 0.7V
   C. 0.5V
   D. 5V

8. The voltage drop across a bridge can be reduced by using _____ rectifiers.
   A. Schottky
   B. high current
   C. Zener
   D. Tunnel

9. The process of smoothing out the DC coming from the rectifier circuit is called:
   A. averaging
   B. filtering
   C. planing
   D. limiting

10. In a capacitor-input filter, current flows from the rectifier circuit:
   A. at all times
   B. only near zero crossings
   C. only during waveform peaks
   D. during the latter part of each half-cycle

11. In a power supply with a bridge rectifier and a capacitor-input filter, the voltage across the capacitor will be roughly:
   A. equal to the transformer output voltage
   B. 1.4X the transformer output voltage
   C. 1.4X the transformer output voltage, minus about 1.4V
   D. equal to the AC line voltage

12. “Ripple” is:
   A. random DC voltage variations
   B. voltage variance with varying loads
   C. a small 60 or 120Hz AC voltage riding on the DC voltage
   D. high frequency glitches on the AC waveform caused by lamp dimmers etc

13. In a power supply using a bridge rectifier, the ripple frequency is:
   A. 60Hz
   B. 120Hz
   C. 180Hz
   D. depends on the transformer

14. It is not worth trying to obtain very low ripple by using huge capacitors because:
   A. big capacitors are expensive
   B. it doesn't work
   C. it results in huge charging current at the peak of each half-cycle
   D. it draws too much power from the AC mains

15. At full load, the DC voltage should drop _____ between the 120Hz capacitor-recharging peaks.
   A. 1-2%
   B. 5-10%
   C. 20-30%
   D. 50-70%

16. The inductor used in an inductor-input filter is called a _____.
   A. ballast
   B. choke
   C. core
   D. winding

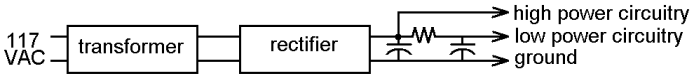
17. The main advantage of a inductor-input filter is:
   A. magnetic radiation is reduced
   B. size and cost are reduced
   C. charging current is not concentrated at waveform peaks
   D. inductor-input filters are safer

18. Two disadvantages of a 60Hz inductor-input filter are (two correct answers):
    A. distortion of the waveform
    B. unreliable operation
    C. large voltage change when load is suddenly changed
    D. high-power filters are very heavy

19. In a swinging choke, when load current increases...
    A. inductance increases
    B. inductance decreases
    C. the resonant frequency increases
    D. the choke resonates with the filter capacitor

20. Two advantages of a 60Hz ferroresonant transformer are (two correct answers):
    A. reduced 60Hz magnetic radiation
    B. near square-wave output
    C. simpler construction
    D. semi-regulated output voltage

21. "Decoupling" as in the two-stage filter shown has the following purpose:



    A. reducing waveform distortion
    B. preventing high-level signals from coupling through the power supply rail
    C. increasing overall gain
    D. improving low frequency response

22. Besides decoupling, the other function of a two-stage power supply filter is:
    A. reducing power supply ripple to low power circuitry
    B. permitting low and high power circuitry to be widely separated
    C. eliminating a ground loop
    D. eliminating DC offset

23. A shunt regulator usually consists of a _____, a _____, and maybe a shunt capacitor.
    A. resistor, tunnel diode,
    B. resistor, zener
    C. resistor, SCR
    D. resistor, inductor

24. A shunt regulator works by:
    A. capacitively filtering the voltage
    B. blocking ripple
    C. dragging down voltage in excess of intended final voltage
    D. separating ripple and routing it to a different wire

25. A series regulator:
    A. is less efficient than a shunt regulator
    B. passes only as much current as the load actually uses
    C. is not suited to low-power circuits
    D. has two terminals

26. In addition to their regulation functions, 78xx and 79xx and similar regulators provide the ollowing additional benefits (two correct answers):
    A. ground loop isolation
    B. overtemperature shutdown
    C. zero-dissipation regulation
    D. short circuit current limit

27. To insure stability of regulator ICs:
    A. connect a choke in series between rectifiers and regulator
    B. connect shunt capacitors at input and output
    C. Earth ground the mounting tab
    D. connect a zener between output and ground

28. A switching regulator operates by varying the:
    A. load current
    B. phase delay
    C. duty cycle
    D. series impedance

29. "Chopping" means:
    A. dividing the sections of a power supply into separate physical units
    B. splitting filter capacitor duties among several capacitors to ease their job
    C. another name for decoupling
    D. rapidly switching a voltage on and off

30. The output of the chopper used in a switching supply is:
    A. a high frequency rectangle wave
    B. a 60Hz square wave
    C. 90° ahead of the power line waveform
    D. unstable

31. This output is turned into smooth DC by a:
    A. a capacitor input filter
    B. a zener regulator
    C. a choke input filter
    D. an incandescent lamp in series

32. Using a variable duty cycle to control output voltage is called:
    A. pulse position modulation
    B. SPSDT switching
    C. pulse width modification
    D. pulse width modulation

33. In a buck converter, the output voltage is _____ the input voltage.
    A. lower than
    B. higher than
    C. equal to
    D. depends on the circuit

34. In a boost converter, the output voltage is _____ the input voltage.
    A. unrelated to
    B. the same as, but higher current than
    C. lower than
    D. higher than

35. The size and weight of a switching power supply can be greatly reduced by:
    A. chopping at a lower frequency
    B. accepting higher ripple
    C. drilling a lot of holes in the case
    D. moving the chopping operation to the primary side of the transformer

36. The main purpose of synchronous rectification is:
    A. smaller size
    B. lower cost
    C. reduced RF radiation
    D. increased efficiency

37. In a bipolar supply:
    A. there are two outputs of the same polarity but different voltages
    B. there are two equal-voltage outputs, one positive, the other negative
    C. there are two unequal-voltage outputs, one positive, the other negative
    D. the input can be positive DC, negative DC or AC

38. An inverter:
    A. turns DC into line-voltage AC
    B. turns +DC into -DC
    C. turns lumpy DC into smooth DC
    D. turns AC at one voltage into AC at another voltage

39. With today's technology, the easiest way to make an efficient true sinewave inverter is:
    A. a huge choke-input filter
    B. a series 180Hz tank
    C. pulse width modulation
    D. class AB amplification

40. In electronics, “UPS” means:
    A. universal power supply
    B. universal power storage
    C. uninterruptible power supply
    D. united power source

41. “VFD” means:
    A. variable filtering drive
    B. variable frequency drive
    C. voltage / frequency drive
    D. variable frequency distribution

CHAPTER 10 - OPTOELECTRONICS

1. “Optoelectronics” is:
    A. any light source used in an electronic device
    B. amplification using light
    C. electronic generation, detection and manipulation of light
    D. wireless communication using light

2.The visible light spectrum is roughly:
    A. 300nm to 150nm
    B. 800nm to 400nm
    C. 1700nm to 800nm
    D. 30μm to 1000nm

3. Infrared light:
    A. has longer wavelengths than visible light
    B. has shorter wavelengths than visible light
    C. is coherent
    D. is collimated

4. In a photoresistive device:
    A. conductance varies inversely with illumination
    B. resistance varies with illumination
    C. resistance varies inversely with illumination
    D. voltage varies with illumination

5. An example of a photoresistive device is:
    A. a photodiode
    B. a phototransistor
    C. A LED
    D. A CdS photocell

6. In a photovoltaic device:
    A. illumination causes a voltage to be created
    B. illumination causes resistance to drop
    C. illumination causes impedance to drop
    D. illumination triggers avalanche mode

7. An example of a photovoltaic device is:
    A. a CdS photocell
    B. A LED
    C. a solar cell
    D. a laser

8. In a photoconductive device:
   A. current flow varies with illumination
   B. resistance varies inversely with illumination
   C. resistance varies directly with illumination
   D. light output varies with current

9. An example of a photoconductive device is:
   A. a LED
   B. a laser
   C. a solar cell
   D. a phototransistor

10. "LDR" means:
   A. LED, red color
   B. light-dependent resistor
   C. light-dependent rectifier (diode)
   D. low-directivity red LED

11. A photodiode can work in _____ or _____ mode.
   A. photoconductive, photovoltaic
   B. photoresistive, photovoltaic
   C. photoresistive, photoconductive
   D. photoconductive, photo-emissive

12. The advantage of a PIN diode over other photodiodes is:
   A. higher sensitivity
   B. lower cost
   C. higher speed
   D. smaller size

13. For a given amount of illumination, a phototransistor exhibits _____ than a photodiode.
   A. greater resistance
   B. greater conductivity
   C. sensitivity over a wider spectrum
   D. lower cost

14. The switching speed of a phototransistor is _____ that of a photodiode with an optimum amplifying circuit.
   A. faster than
   B. the same as
   C. unpredictable compared to
   D. slower than

15. The forward voltage drop of a red LED is around ____ volts.
   A. 1
   B. 1.5
   C. 3-4
   D. 2

16. The forward voltage drop of a blue LED is around ___ volts.
   A. 2
   B. 3-4
   C. 6.3
   D. 8

17. True or false: carefully regulating the voltage across a red LED is an effective way of controlling current.
   A. true
   B. false

18. White LEDs do not have an unlimited life expectancy because:
   A. they operate at high temperatures
   B. they use a phosphor which deteriorates
   C. their red, green and blue components interfere
   D. they cannot be made free of impurities

19. A laser emits light at a single wavelength. This is called _____ light.
   A. coordinated
   B. coherent
   C. collimated
   D. concentrated

20. An optocoupler pairs a _____ and a _____.
   A. laser, photodetector
   B. laser, LED
   C. LED, photodetector
   D. photodiode, phototransistor

21. Another name for optocoupler is:
   A. optoresistor
   B. opto-isolator
   C. optodiode
   D. light dependent FET

22. True or false: fiber optics can transmit any wavelength of infrared or visible light efficiently over long distances.
   A. true
   B. false

ANSWERS

Chapter 1, Analog Signal Basics

1-C 2-C 3-C 4–A 5-D 6-B 7-C 8-B 9-A 10-B
11-C 12-B 13-C 14-B 15-C 16-B 17-B 18-D
19-A 20-C 21-D 22-A 23-C 24-B 25-B 26-C
27-B 28-A 29-B

Chapter 2, Active Devices

1-D 2-A 3-A 4-C 5-B 6-D 7-D 8-D 9-D 10-B,C
11-C 12-B 13-D 14-B 15-B 16-C 17-A 18-D 19-A
20-B 21-C 22-B 23-B 24-C 25-B 26-D 27-C 28-A
29-A

Chapter 3, Transistor Amplifiers

1-B 2-D 3-C 4-B 5-B 6-B 7-C 8-A 9-B 10-A
11-A 12-B 13-C 14-D 15-B 16-D 17-B 18-D 19-B
20-B 21-A 22-C 23-D 24-A 25-B 26-A 27-C 28-B
29-D 30-A 31-C 32-A 33-C 34-A 35-C 36-B 37-A
38-C 39-A 40-C 41-D 42-A 43-D

Chapter 4, Operational Amplifiers

1-D 2-A 3-B 4-D 5-A 6-C 7-B 8-A 9-C 10-A
11-D

Chapter 5, Analog Integrated Circuits

1-D 2-A 3-C 4-B 5-A 6-D 7-D 8-B 9-A 10-B
11-C 12-D 13-B 14-B 15-A 16-C 17-B 18-D
19-A 20-B 21-B 22-B 23-C 24-B 25-C 26-B
27-D 28-B

Chapter 6, Oscillators

1-B 2-C 3-A 4-C 5-B 6-D 7-A 8-C 9-B 10-C
11-D 12-B 13-B 14-A 15-C 16-A 17-B 18-B
19-A 20-C 21-B 22-B 23-C 24-A 25-A 26-B
27-D 28-B 29-D

Chapter 7, Wave Shaping Circuits

1-B 2-B 3-B 4-B 5-A 6-B 7-D 8-C 9-B 10-A
11-C 12-C 13-C 14-A 15-B 16-C 17-B 18-B

Chapter 8, Filters

1-C 2-A 3-A 4-A 5-C 6-B 7-A 8-B 9-A 10-B
11-C 12-A 13-A 14-D 15-B 16-B 17-C 18-B
19-C 20-C 21-B 22-C 23-B 24-D 25-A 26-B
27-B 28-A 29-D 30-B,D 31-A,D 32-C 33-B
34-A 35-B 36-D 37-B 38-B 39-A 40-B 41-C
42-B

Chapter 9, Power Supplies

1-B 2-B 3- A 4-B 5-C 6-B 7-A 8-A 9-B 10-C
11-C 12-C 13-B 14-C 15-B 16-B 17-C
18-C,D 19-B 20-B,D 21-B 22-A 23-B 24-C
25-B 26-B,D 27-B 28-C 29-D 30-A 31-C
32-D 33-A 34-D 35-D 36-D 37-B 38-A 39-C
40-C 41-B

Chapter 10, Optoelectronics

1-C 2-B 3-A 4-C 5-D 6-A 7-C 8-A 9-D 10-B
11-A 12-C 13-B 14-D 15-D 16-B 17-B 18-B
19-B 20-C 21-B 22-B

End of EM3 practice exam